Jurnal Teknik Informatika (JUTIF)
Vol. 5, No. 4, August 2024, pp. 579-591

DOI: https://doi.org/10.52436/1.jutif.2024.5.4.2217
p-ISSN: 2723-3863
e-ISSN: 2723-3871

SVM OPTIMIZATION WITH INFORMATION GAIN FEATURE SELECTION TO INCREASE THE ACCURACY OF SENTIMENT ANALYSIS OF INCREASING THE COST OF THE HAJJ

Manarul Hidayat[*1], Arief Wibowo[2]

[1,2]Master of Computer Science, Faculty of Information Technology, Universitas Budi Luhur, Indonesia
Email: [1]manarulhidayat.id@gmail.com, [2]arief.wibowo@budiluhur.ac.id

(Article received: June 5, 2024; Revision: June 15, 2024; published: August 21, 2024)

Abstract

Everyone's freedom to express their opinions is now poured into a platform known as social media. This platform allows people in the digital world to communicate with each other using the internet. YouTube is one of the most popular social media platforms worldwide. In 2023, the Government, in this case the Ministry of Religious Affairs of the Republic of Indonesia and Commission VIII of the House of Representatives have approved the Hajj Travel Cost 1444 H/2023 AD with a range of Rp90,050,637.26 per regular pilgrim. In contrast to the government of the Kingdom of Saudi Arabia, which implemented a policy of reducing the cost of the Hajj package by 30% from 2022. This has caused pros and cons to the hajj cost increase. Public opinion on social media is the focus of this research to conduct sentiment analysis. Sentiment analysis has been developed through various methods, but there are still many challenges to produce accurate sentiment analysis. The challenges include accuracy, binary classification, data sparsity, and polarity shift. One of the challenges in improving accuracy is the focus of this research. In this study, the Support Vector Machine method is applied and Information Gain feature selection is added. The accuracy results obtained in this study are the Support Vector Machine method (87%) and Support Vector Machine combine with information gain feature selection (89%). It can be concluded, the support vector machine method combined with information gain feature selection proves an increase in accuracy by 2%.

Keywords: *Hajj Cost Increase, Information Gain, Sentiment Analysis, Support Vector Machine, Youtube.*

OPTIMASI SVM DENGAN SELEKSI FITUR INFORMATION GAIN UNTUK PENINGKATAN AKURASI ANALISIS SENTIMEN KENAIKAN BIAYA IBADAH HAJI

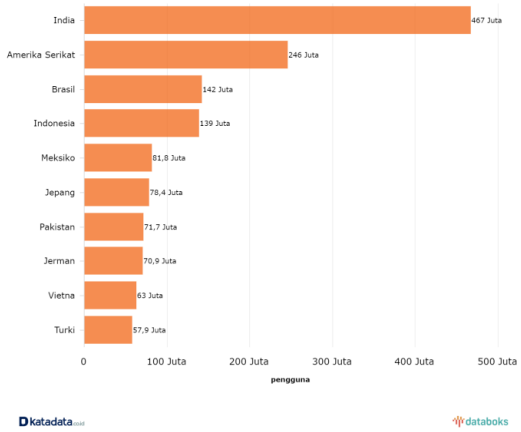
Abstrak

Kebebasan setiap orang untuk menyampaikan pendapatnya kini dituangkan ke dalam platform yang dikenal dengan media sosial. Platform ini memungkinkan orang-orang di dunia digital untuk berkomunikasi satu sama lain dengan menggunakan internet. *YouTube* adalah salah satu *platform* media sosial terpopuler di seluruh dunia. Pada tahun 2023 Pemerintah dalam hal ini Kementerian Agama Republik Indonesia dan Komisi VIII Dewan Perwakilan Rakyat telah menyetujui Biaya Perjalanan Ibadah Haji (Bipih) 1444 H/2023 M dengan kisaran sebesar Rp90.050.637,26 per jemaah haji reguler. Berbeda dengan pemerintah Kerajaan Arab Saudi yang menerapkan kebijakan menurunkan biaya paket haji sebesar 30% dari tahun 2022. Hal ini menyebabkan pro dan kontra terhadap kenaikan biaya haji tersebut. Opini masyarakat di media sosial tersebut menjadi fokus penelitian ini untuk dilakukan analisis sentimen. Analisis sentimen telah dikembangkan melalui berbagai metode, namun masih banyak tantangan untuk menghasilkan analisis sentimen yang akurat. Tantangan yang dimaksud seperti akurasi, klasifikasi binar, *data sparsity*, dan *polarity shift*. Salah satu tantangan dalam meningkatkan akurasi menjadi fokus dalam penelitian ini. Pada penelitian ini, metode *Support Vector Machine* diterapkan dan ditambahkan fitur seleksi *Information Gain*. Hasil akurasi yang diperoleh pada penelitian ini yaitu metode *Support Vector Machine* (87%) dan *Support Vector Machine* yang dikombinasikan dengan seleksi fitur *information gain* (89%). Dapat disimpulkan, metode *support vector machine* yang dikombinasikan dengan seleksi *fitur information gain* membuktikan terjadi peningkatan akurasi sebesar 2%.

Kata kunci: Analisis Sentimen, Information Gain, Kenaikan Biaya Haji, Support Vector Machine, Youtube.

1. PENDAHULUAN

Kebebasan setiap orang untuk menyampaikan pendapatnya kini dituangkan ke dalam platform yang dikenal dengan media sosial. *Platform* ini memungkinkan orang-orang di dunia digital untuk berkomunikasi satu sama lain dengan menggunakan internet. *YouTube* adalah salah satu *platform* media sosial terpopuler di seluruh dunia [1]. Media sosial dengan jumlah pengguna aktif terbanyak global dapat dilihat pada Gambar 1 [2]



Gambar 1. Media Sosial dengan Jumlah Pengguna Aktif Terbanyak Global

Pada tahun 2023 Pemerintah dalam hal ini Kementerian Agama Republik Indonesia dan Komisi VIII Dewan Perwakilan Rakyat telah menyetujui Biaya Perjalanan Ibadah Haji (Bipih) 1444 H/2023 M dengan kisaran sebesar Rp90.050.637,26 per jemaah haji reguler. Angka ini terdiri dari dua komponen, yaitu Biaya Perjalanan Ibadah Haji (Bipih) yang ditanggung jemaah sebesar Rp49.812.700,26 (55,3%), dan penggunaan nilai manfaat per jemaah sebesar Rp40.237.937 (44,7%).

Penetapan biaya perjalanan haji dikukuhkan dengan Keputusan Menteri Agama Nomor 352 Tahun 2023 mengenai biaya perjalanan haji reguler tahun 1444H/2023 M dan penggunaan nilai manfaat serta Keputusan Presiden Nomor 7 Tahun 2023 [3]. Berbeda dengan pemerintah Kerajaan Arab Saudi yang menerapkan kebijakan menurunkan biaya paket haji sebesar 30% dari tahun 2022, tidak ada batasan usia bagi jemaah bahkan mendapat tambahan kuota sebanyak 221.000 jemaah untuk Indonesia. Hal ini menyebabkan pro dan kontra terhadap kenaikan biaya ibadah haji tersebut. Pembahasan mengenai kenaikan biaya ibadah haji tahun 2023 menjadi sumber data dalam penelitian ini. Opini masyarakat di media sosial tersebut menjadi fokus penelitian ini untuk dilakukan analisis sentimen.

Analisis sentimen adalah sebuah sistem bertujuan untuk mengidentifikasi sikap dan kepuasan pengguna terhadap topik atau masalah tertentu dengan melihat opini mereka di media sosial dan komentar mereka di blog dan website [4]. Analisis sentimen sendiri adalah metode ekstraksi data teks untuk mengetahui apakah sentimen memiliki nilai positif atau negatif [5]. Analisis sentimen telah dikembangkan dengan menggunakan berbagai metode, namun terdapat tantangan untuk menghasilkan analisis sentimen yang akurat. Tantangan yang dimaksud seperti akurasi, klasifikasi binar, *data sparsity*, dan *polarity shift* [6]. Salah satu tantangan untuk meningkatkan akurasi menjadi fokus dalam penelitian ini.

Klasifikasi analisis sentimen dapat dilakukan dengan berbagai teknik; namun, metode yang paling umum digunakan adalah antara lain *Naïve Bayes* (NB), *Logistic Regression* (LR), *Multinomial Naïve Bayes* (MNB), *Random Forest* (RF), *Maximum Entropy* (MaxEnt), *Conditional Random Field* (CRF), dan *Support Vector Machines* (SVM) [7]. Sejumlah penelitian telah dilakukan mengenai analisis sentimen kenaikan harga, antara lain: Rudy Asrianto dan Melda Herwinanda membahas analisis sentimen terkait kenaikan harga kebutuhan pokok di media sosial *YouTube* menggunakan *support vector machine* [1]. Kemudian penelitian yang dilakukan oleh Dani Juhaeni dan Arief Wibowo membahas penerapan metode *Naïve Bayes* terkait Wacana Kenaikan Harga Tiket Candi Borobudur Pada *Twitter* [8]. Kemudian penelitian yang dilakukan oleh Siti Nurhaliza dkk terkait Klasifikasi Sentimen Masyarakat di *Twitter* Terhadap Kenaikan Harga BBM dengan Metode *Support Vector Machine* [9]. Kemudian penelitian yang dilakukan oleh Rifna Savira dkk membahas Analisis Sentimen Pada *Twitter* terkait Kenaikan BBM 2022 Dengan *Lexicon* dan *Support Vector Machine* [10].

Dari beberapa algortima yang paling banyak digunakan untuk klasifikasi data yaitu algortima *Support Vector Machine*. *Support Vector Machine* merupakan metode *supervised learning* yang menganalisa data mengidentifikasi pola yang digunakan untuk klasifikasi [11]. Savira, dkk (2023) melakukan penelitian menggunakan metode *Support Vector Machine*, hasil akurasi yang diperoleh 79%. Hasil akurasi penelitian yang sama menggunakan metode *Support Vector Machine* oleh Utama dkk (2019) yaitu 78.18%. Penelitian Ilmawan dan Mude (2020) membandingkan kinerja dari sentimen analisis yang dilakukan antara Naïve Bayes dan SVM. Hasil dari penelitian tersebut, metode SVM menghasilkan akurasi tertinggi sebesar 81.46%, dibandingkan metode Naïve Bayes mendapatkan hasil akurasi sebesar 75.41%. Oleh sebab itu, penelitian ini menerapkan metode *SupportVector Machine*. Metode *SVM* adalah metode klasifikasi yang baik dan bersifat regresi. Kelebihan dari metode *Support Vector Machine* yaitu dapat menyelesaikan penelitian bersifat linier atau non-linear [13]. Namun *Support Vector Machine* masih memiliki kekurangan yaitu dalam hal komputasi data dengan jumlah yang besar [14]. Jumlah data yang diolah mempengaruhi nilai akurasi yang diperoleh [15]. Guna mengatasi permasalahan

tersebut, diperlukan sebuah algoritma pemilihan fitur yang dapat meningkatkan kinerja *support vector machine* dalam mengklasifikasikan analisis sentiment [16].

Pemilihan fitur merupakan langkah penting dalam mempengaruhi kinerja klasifikasi. Banyak penelitian mencoba menambahkan algoritma pengoptimalan untuk menyajikan metode seleksi untuk meningkatkan efeknya. Seleksi fitur digunakan untuk mengurangi jumlah set fitur yang besar menjadi lebih kecil sehingga meningkatkan nilai akurasi klasifikasi.

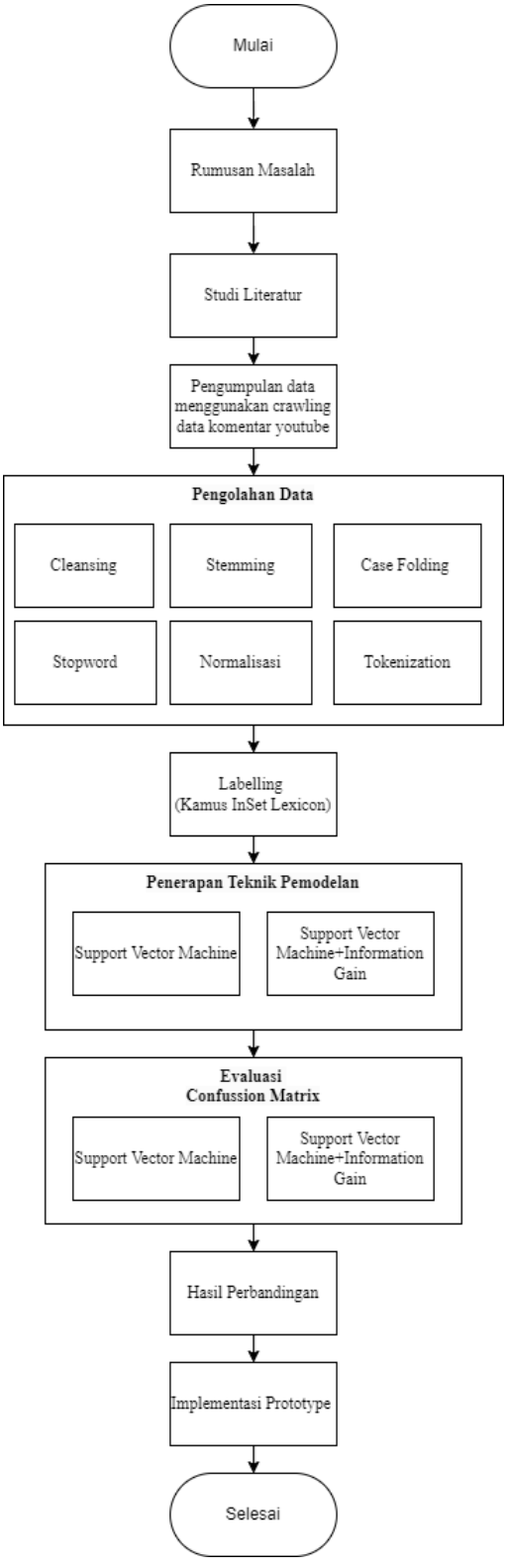
Dalam proses fitur seleksi, *information gain* merupakan metode pemilihan fitur terbaik [17]. Menurut hasil perbandingan untuk algoritma seleksi fitur, antara *Chi Square*, *Information Gain*, *Backward Elimination*, dan *Forward Selection*, didapatkan bahwa *Information Gain* dianggap sebagai algoritma seleksi fitur terbaik dengan nilai akurasi terbesar dengan selisih 0,7% dari *Chi Square*, selisih 2,25% dari *Backward Elimination*, dan selisih 13,83% dari *Forward Selection* [18]. Dalam Negara dkk., (2020), Uysal & Gunal (2012), *Information Gain* sering lebih unggul daripada yang lainnya. *Informasi Gain* mengukur jumlah informasi yang ada dan tidak ada pada suatu kata, ini penting penting untuk membuat keputusan klasifikasi yang tepat untuk kelas apapun. *Information Gain* adalah salah satu metode *filter* yang berhasil dalan pengklasifikasian teks. Dengan demikian, fitur seleksi yang digunakan dalam penelitian ini adalah fitur seleksi *information gain* [19].

2. METODE PENELITIAN

Proses pada penelitian ini terdiri dari beberapa proses, seperti yang ditunjukkan pada gambar 2.

Berikut ini merupakan penjelasan merupakan penjelasan tahapan penelitian pada Gambar 2.

1. Rumusan Masalah
   Pada tahap ini, perumusan masalah dilakukan berdasarkan latar belakang masalah yang terjadi sehingga beberapa masalah dapat diselesaikan dengan metode yang sesuai dengan kebutuhan pengguna dan sistem.
2. Studi Literatur
   Tahapan ini dilakukan dengan mempelajari literatur untuk memperoleh uraian tentang teori yang berkaitan dengan sistem informasi, teknik data mining dan pengujian sistem. Selain itu, pada tahap ini dilakukan penelitian dan analisis jurnal dan tesis tentang *data mining*.
3. Pengumpulan Data
   Pada tahap ini dilakukan pengumpulan data yang diambil dari media sosial *youtube* berupa komentar yang diambil dari URL *youtube* dari beberapa media atau stasiun televisi nasional yang membahas terkait isu kenaikan biaya ibadah haji tahun 2023.



Gambar 2. Tahapan Penelitian

4. Pengolahan Data
   Pada tahap ini data mentah berupa komentar diolah menjadi data yang siap di proses, sehingga diperlukan proses *cleansing*, membuang imbuhan *(stemming)*, mengubah huruf kapital menjadi huruf kecil *(case folding)*, pemisahan kata yang tidak memiliki makna *(stopword)*,

proses menormalkan semua spasi (normalisasi), dan pemotongan kata *(tokenizing).*

5. *Labelling*
   Proses *labelling* sentimen menggunakan kamus opini *InSet Lexicon*. Apabila kata dalam komentar ditemukan dalam kamus maka kata tersebut diakumulasikan nilai *polarity score* sehingga dapat menentukan label positif atau negatif.
6. Penerapan Teknik Pemodelan
   Pada tahap ini dilakukan penerapan teknik pemodelan data mining. Pemodelan yang akan diimplementasi yaitu SVM dan SVM+*Information Gain*.
7. Evaluasi menggunakan *confussion matrix*
   Pada tahap ini dilakukan pengujian model untuk melihat akurasi terbaik diantara model yang diterapkan.
8. Hasil Perbandingan
   Selanjutnya dilihat hasil akurasi terbaik yang diperoleh dari model yang diterapkan.
9. Implementasi Prototipe
   Pada tahap ini pemodelan yang telah diterapkan selanjutnya dirancang kedalam program.

2.1 *Data Mining*

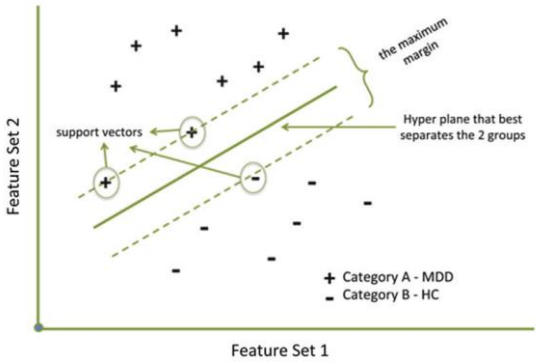
Data mining merupakan pengumpulan data atau analisis data melalui pola tersembunyi, hubungan antar elemen atau pembuatan model melalui analisis matematika. Tujuan *data mining* yaitu menemukan pola yang tidak diketahui sebelumnya untuk menyelesaikan sebuah permasalahan dan mendapatkan informasi menggunakan metode *data mining*. Untuk menghasilkan model yang baik maka data atau informasi yang dilihat dari cara pengambilan data sebelumnya sehingga dapat diprediksi di masa depan. Oleh sebab itu, semakin lampau data yang diperoleh maka semakin akurat hasil yang diperoleh [20].

2.2 Analisis Sentimen

Analisis sentimen merupakan proses menentukan sentimen dan mengklasifikasikan polaritas tekstual suatu dokumen atau bagian untuk menentukan kategori sebagai sentimen positif, negatif, atau netral. Analisis sentimen juga dapat dibandingkan dengan penambangan opini karena berfokus pada pendapat, yaitu pendapat yang disampaikan secara positif atau negatif. Analisis sentimen melibatkan penambangan data, yang menganalisis dan mengekstrak data teks dalam entitas seperti layanan, produk, orang, dan topik [21].

2.3 Support Vector Machine

Salah satu metode populer untuk klasfikasi mapun regresi yaitu *Support Vector Machine* (SVM). SVM termasuk dalam *supervised learning* yang artinya dalam implementasi diperlukan tahap training dan tahap testing [22]. *Support Vector Machine* memaksimalkan jarak (margin) antar dua kelas untuk mendapatkan hyperplane terbaik [23]. Pencarian *hyperlane* dapat dilihat pada gambar 3.



Gambar 3. Pencarian *Hyperline*

Dalam menentukan *hyperplane* data dapat mempertimbangkan titik data menggunakan {(x_1, y_1), (x_2, y_2), (x_3, y_3),…, ((x_n, y_n)} dimana y_n = 1/-1 sebuah konstanta yang menunjukan kelas yang titik x_n miliki dan n = jumlah sampel, dapat didefinisikan persamaan [24]:

$$w.\,x + b = 0 \quad (1)$$

Dimana b adalah scalar dan w adalah vector p-dimensi. Pada kelas positif dapat diinisiasi dengan y_n=1, sehingga dapat didefinisikan bidang pembatas pertama dengan persamaan [24][24]:

$$w.\,x + b = 1 \quad (2)$$

Sedangkan pada kelas negatif dapat diinisisasi dengan y_n=-1, sehingga dapat didefinisikan bidang pembatas kedua dengan persamaan [24]:

$$w.\,x + b = -1 \quad (3)$$

Jika data dipisahkan secara linear kita dapat memaksimalkan jarak mereka dengan $2|w|$ atau dengan meminimkan $|w|2$. Maka pencarian *hyperplane* optimal dengan memaksimalkan kedua bidang pembatas dapat digambarkan sebagai persamaan [24]:

$$y_i(w.\,x_i + b) \geq 1 \quad (4)$$

Dalam hal teknik klasifikasi *support vector machine* dapat menggunakan data linear atau non-linear. Untuk menyeselesaikan permasalahan non-linear, algorima support vector machine memanfaatkan fungsi kernel. Namun, penggunaan fungsi kernel dapat menyebabkan over fitting, oleh karena itu digunakan soft margin dan konstanta C untuk mengontol antara margin dan error. Fungsi kernel yang digunakan yaitu Linear dan Radial Basis Function (RBF) dengan persamaan kernel [24]:

Persamaan kernel Linear:

$$K(x_t, x_{tn}) = x_t \quad (5)$$

Persamaan kernel *Radial Basis Function*:

$$K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right), \gamma > 0 \quad 6)$$

2.4 Information Gain

Information Gain adalah salah satu metode seleksi fitur yang paling penting dan paling populer. Metode *Information Gain* digunakan untuk memilih fitur penting sehubungan dengan atribut kelas untuk mengurangi dimensi untuk klasifikasi yang efisien dan untuk menyatakan jumlah informasi yang diberikan oleh fitur kata ke seluruh kategori. Jika jumlah Information Gain lebih besar, maka kata itu yang bisa dipilih [25]. Rumus Formula *Information Gain* sebagai berikut [25]:

$$GD(D,t) = \sum_{i=1}^{m} P(c_i)\, log\, P(c_i) + \quad (7)$$

$$P(t) \sum_{i=1}^{m} P(C_i \mid t)\, log\, P\,(C_i \mid t) + \quad (8)$$

$$P(\bar{t}) \sum_{i=1}^{m} P(c_i \mid \bar{t})\, log\, P\,(C_i \mid \bar{t}) + \quad (9)$$

Dalam rumus $p(c_1)$ mewakili probabilitas bahwa sampel arbitrer termasuk dalam kategori (c_1), $p(t)$ mewakili kemungkinan istilah tersebut ditemukan dalam teks, $p(c_1 \mid t)$menunjukkan probabilitas bahwa sebuah teks termasuk dalam kategori (c_1) sedangkan $p(t)$ muncul dalam teks ini, menunjukkan probabilitas bahwa istilah t tidak ada dalam teks, $p(c_1 \mid t)$ mewakili probabilitas bahwa teks termasuk dalam kategori c_1 sedangkan t tidak muncul dalam teks ini., $p(c_1 \mid t)$ mewakili jumlah total kategori Ternyata, nilai IG (t) yang lebih besar menunjukkan lebih banyak informasi yang diberikan oleh kata fitur ke kategori tersebut, dan semakin penting kata fitur tersebut [26].

2.5 Lexicon

Lexicon-based approach merupakan metode ilmiah yang umum digunakan dalam penelitian analisis sentimen. Metode ini bekerja dengan menggunakan kamus kata atau corpus dengan bobot setiap kata sebagai sumber bahasa atau kosa kata. Hasil analisis menggunakan metode ini berupa klasifikasi sentimen positif, negatif, dan netral. Teknik ini merupakan komponen dari *machine learning*. Kualitas hasil bergantung pada kamus kata atau corpus yang digunakan [27].

Salah satu kamus (leksikal) adalah *Inset lexicon* atau *Indonesian Sentiment Lexicon*, yaitu *Lexicon Based* yang menggunakan kamus kata-kata berbahasa Indonesia, penggunaan *Inset lexicon* dapat menyederhanakan ekstraksi data tekstual karena teks tersebut tidak perlu diterjemahkan ke dalam Bahasa Inggris [28]. *Inset lexicon* memiliki kinerja dan performansi yang cukup memuaskan sebagai kamus sentimen Bahasa Indonesia dengan tingkat akurasi sebesar 65.78% [29].

Kamus *lexicon* ini terdiri dari 3.609 kata positif dan 6.609 kata negatif. Bobot pada kamus *lexicon* ini memiliki nilai dengan rentang skor dari -5 sampai +5 berbahasa Indonesia yang telah memiliki bobot nilai atau *polarity score* pada setiap katanya dengan kisaran bobot antara -5 sampai +5. Contoh kamus *lexicon* yang memiliki kata positif dan negatif disertai skor tiap kata dijabarkan pada tabel 1.

Tabel 1. Contoh Kamus InSet *(Indonesia Sentiment)* Lexicon

No	Negatif		Positif	
	Kata	Skor	Kata	Skor
1	azab	-5	paripurna	1
2	zalim	-5	terimakasih	5
3	gendeng	-2	penerangan	2
4	mengkhianat	-4	welas	4
5	kebodohan	-3	belas	2
6	gatau	-1	allah	5
7	anjir	-3	mohon	2
8	koit	-5	merespon	3
9	baper	-4	suka cita	4
10	kalap	-3	terawat	5

2.6 Confusion Matrix

Confusion matrix adalah salah satu instrumen untuk mengukur klasifikasi biner dua kelas melalui *matrix* berbentuk tabel. Melalui *Confusion matrix*, maka klasifikasi dapat diukur kinerjanya. Hasil klasifikasi berupa *True Positive* dan *True Negative* artinya klasifikasi benar, serta *False Positive* dan *False Negative* artinya klasifikasi salah [31]. *Confusion Matrix* menjelaskan ukuran n x n yang berhubungan dengan pengklasifikasi yang mewakili klasifikasi yang diprediksi dan aktual, di mana n adalah jumlah kelas yang berbeda [30].

Tabel 1 merupakan contoh matrix dari *Confusion matrix*.

Tabel 2. *Confusion Matrix* Dua Kelas

Predicted Class	True Class	
	Positive	Negative
Positive	TP (True Positive)	FP (False Positive)
Negative	FN (False Negative)	TN (True Negative)

Hasil dari *Confusion matrix* berupa satuan persen dan digunakan untuk menghasilkan nilai *precision*, *recall*, dan *accuracy*. *Precision* adalah tingkat keakuratan antara informasi yang diminta oleh pengguna dengan respon yang diberikan oleh sistem. *Recall* adalah tingkat keberhasilan dalam menemukan kembali sebuah informasi. Sedangkan *accuracy* adalah tingkat kedekatan antara nilai prediksi dengan nilai aktual. Perhitungan *precision*, *recall*, dan *accuracy* dijelaskan sebagai berikut [32]:

1. *Precision*

$$\frac{TP}{TP + FP} x100\% \quad (10)$$

2. *Recall*

$$\frac{TP}{TP + FN} x100\% \quad (11)$$

3. *Accuracy*

$$\frac{TP + TN}{Total\ Sample} x100\% \quad (12)$$

Keterangan:

TP = True Positive, artinya data yang digunakan merupakan data yang benar mengikuti kegiatan dengan kriteria yang ditentukan

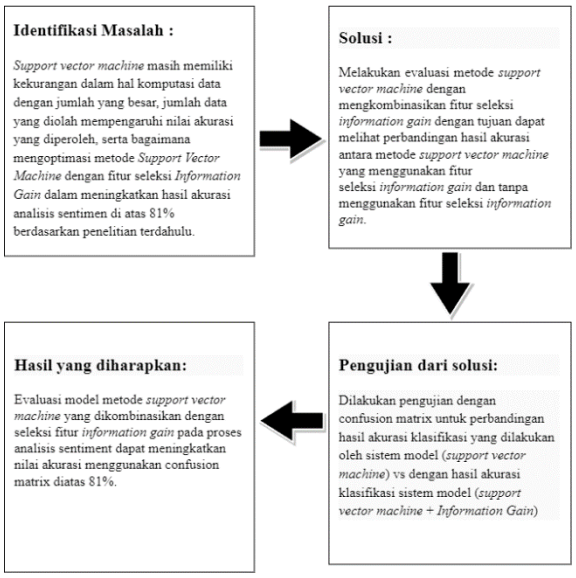
FP = False Positive, artinya data yang digunakan merupakan data yang benar mengikuti kegiatan namun nilai tidak memenuhi kriteria yang ditentukan.

FN = False Negative, artinya data yang digunakan merupakan bukan data yang benar namun nilai yang didapat sesuai dengan kriteria yang ditentukan.

TN = True Negative, artinya data yang digunakan merupakan bukan data yang benar dan nilai yang didapat tidak memenuhi kriteria yang ditentukan

3. HASIL DAN PEMBAHASAN

Bab ini menjelaskan langkah-langkah ekstraksi data yang digunakan untuk mengklasifikasi kata-kata sebagai kata positif dan kata negatif. Studi kasus dalam penelitian ini menggunakan data yang diambil adalah komentar *youtube* dari beberapa media atau stasiun televisi nasional yang membahas terkait isu kenaikan biaya ibadah haji tahun 2023. Dengan kerangka konsep sebagai mana pada Gambar 4 berikut:



Gambar 4 Kerangka Konsep Penelitian

3.1. Pengumpulan Data

Pengumpulan data dilakukan dengan cara mengumpulkan data dari komentar *YouTube* yang diambil dari URL *youtube* dari beberapa media atau stasiun televisi nasional yang membahas terkait isu kenaikan biaya ibadah haji tahun 2023. Data diambil dari 7 Agustus 2023 hingga 18 Agustus 2023 sehingga terkumpul sebanyak 2.500 dataset.

Tabel 3. Data

No	komentar
1.	salam mabrur.....semoga ada penurunan biar jamaah yang berangkat tahun ini tidak kaget bisa melunasi dan berangkat dengan lega.....sadarlah banyak yang ekonomi pas-pasan.....
2.	semestinya negara ini bersukur atas kesadarannya rakyatnya patuh dan taat kepada allah penciptanya. bila semua manusia patuh terhadap aturan penciptanya dalam alquran dijanjikan negeri ini subur dan makmur serta tidak ada perpecahan di muka bumi ini. mestinya negara mendukung dan menfasilitasi dan meringankan beban jamaah haji. tidak mencari keuntungan apa lagi memberatkan dengan biaya-biaya yang mesti tidak perlu.
3.	harusnya kalau mau dinaikan diberlakukannya bagi yang baru daftar jangan yang ssudah daftar apalagi yang ssudah mau berangkat
4.	sebaiknya dpr jangan hanya membahas masalah kenaikan ongkos haji tahun 2023 saja tapi bahas sekaligus untuk ongkos haji tahun " berikutnya agar hal hal semacam ini tidak terulang lagi, karena selalu membuat kegaduhan ,keresahan dan memberatkan calon jemaah haji .
5.	kecerdasan menteri agama membawa kemudaratan bagi umat muslim di indonesia.....
6.	YANG menaikkan BIAYA HAJI SEMOGA ALLAH SWT MENGAZABNYA SANGAT PEDIH ,
7.	pihak saudi arab sangat bijak mempermsudah sarana ibadah haji. hanya di indonesia yang di jadikan lahan bisnis oleh pemerintah. ini bukan politik. tapi ini soal ibadah umat islam.
8.	bulan april ini saya dan keluarga akan tunai umrah., semoga semuanya lancar serta dipermsudahkan segala urusan umrah nanti..
9.	ssudah tertulis,,,,apabila pekerjaan/jabatan tidak sesuai kemampuan tunggu kehancurannya...
10.	berhitungnya jugan dibanding DENGAN biaya umroh yang 9 hari x â± 30 juta naik pesawat 1x pulang pergi / jadi haji yang 40 hari harus 120 juta, emang berhaji naik pesawatnya 4x pp , kan sama naik pesawatnya 1xpp, yang berbeda itu biaya makan /hotel dan selama di arofah muzdalifah dan mina

Pada Tabel 2 diketahui data yang diperoleh masih lengkap dengan beragam karakter, sehingga perlu diproses untuk menghilangkan karakter yang tidak diperlukan. Tujuan dari proses ini adalah agar data yang diproses menjadi lebih bersih dan siap diproses pada tahap selanjutnya.

3.1.1 *Preprocessing*

Tahap pengolahan data asli yang sudah diperoleh berupa data *text* dari sentimen kenaikan biaya ibadah haji. *Preprocessing* bertujuan untuk mengurangi gangguan, memperjelas ciri-ciri, serta menyesuaikan data asli dengan kebutuhan. Berikut ini adalah tahapan dalam *preprocessing*:

a. *Cleansing*

Pada tahap ini, data akan disaring untuk menghapus karakter seperti tanda baca dan karakter

khusus ("!"#$%&()*+-./:;<=>?@[]^_`{|}~\n,). *Cleansing* bertujuan untuk membersihkan karakter yang tidak diperlukan untuk mengurangi *noise*. Hasil data-data yang telah memalalui tahap *cleansing* sebagaimana Tabel 3 berikut:

Tabel 4. *Cleansing* Dokumen

No	Kalimat
1.	salam mabrur semoga ada penurunan biar jamaah yang berangkat tahun ini tidak kaget bisa melunasi dan berangkat dengan lega sadarlah banyak yang ekonomi pas pasan
2.	semestinya negara ini bersukur atas kesadarannya rakyatnya patuh dan taat kepada allah penciptanya bila semua manusia patuh terhadap aturan penciptanya dalam alquran dijanjikan negeri ini subur dan makmur serta tidak ada perpecahan di muka bumi ini mestinya negara mendukung dan menfasilitasi dan meringankan beban jamaah haji tidak mencari keuntungan apa lagi memberatkan dengan biaya biaya yang mesti tidak perlu
3.	harusnya kalau mau dinaikan diberlakukannya bagi yang baru daftar jangan yang ssudah daftar apalagi yang ssudah mau berangkat
4.	sebaiknya dpr jangan hanya membahas masalah kenaikan ongkos haji tahun saja tapi bahas sekaligus untuk ongkos haji tahun berikutnya agar hal hal semacam ini tidak terulang lagi karena selalu membuat kegaduhan keresahan dan memberatkan calon jemaah haji
5.	kecerdasan menteri agama membawa kemudaratan bagi umat muslim di indonesia
6.	YANG menaikkan BIAYA HAJI SEMOGA ALLAH SWT MENGAZABNYA SANGAT PEDIH ,
7.	pihak saudi arab sangat bijak mempermsudah sarana ibadah haji hanya di indonesia yang di jadikan lahan bisnis oleh pemerintah ini bukan politik tapi ini soal ibadah umat islam
8.	bulan april ini saya dan keluarga akan tunai umrah semoga semuanya lancar serta dipermsudahkan segala urusan umrah nanti
9.	ssudah tertulis apabila pekerjaan jabatan tidak sesuai kemampuan tunggu kehancurannya
10.	berhitungnya jugan dibanding DENGAN biaya umroh yang hari x â juta naik pesawat pulang pergi jadi haji yang hari harus juta emang berhaji naik pesawatnya pp kan sama naik pesawatnya yang berbeda itu biaya makan hotel dan selama di arofah muzdalifah dan mina

b. *Case Folding*

Pada tahap ini ini mencakup standarisasi huruf dalam dokumen, huruf kapital akan diubah menjadi huruf kecil, dan semua karakter yang tidak termasuk huruf dihapus karena dianggap sebagai pembatas. Data dalam dokumen yang telah diproses *case folding* sehingga huru-huruf diseragamkan dapat dilihat pada Tabel 4.

Tabel 5. Proses *Case Folding*

No	Kalimat
1.	salam mabrur semoga ada penurunan biar jamaah yang berangkat tahun ini tidak kaget bisa melunasi dan berangkat dengan lega sadarlah banyak yang ekonomi pas pasan
2.	semestinya negara ini bersukur atas kesadarannya rakyatnya patuh dan taat kepada allah penciptanya bila semua manusia patuh terhadap aturan penciptanya dalam alquran dijanjikan negeri ini subur dan makmur serta tidak ada perpecahan di muka bumi ini mestinya negara mendukung dan menfasilitasi dan meringankan beban jamaah haji tidak mencari keuntungan apa lagi memberatkan dengan biaya biaya yang mesti tidak perlu
3.	harusnya kalau mau dinaikan diberlakukannya bagi yang baru daftar jangan yang ssudah daftar apalagi yang ssudah mau berangkat
4.	sebaiknya dpr jangan hanya membahas masalah kenaikan ongkos haji tahun saja tapi bahas sekaligus untuk ongkos haji tahun berikutnya agar hal hal semacam ini tidak terulang lagi karena selalu membuat kegaduhan keresahan dan memberatkan calon jemaah haji
5.	kecerdasan menteri agama membawa kemudaratan bagi umat muslim di indonesia
6.	yang menaikkan biaya haji semoga allah swt mengazabnya sangat pedih
7.	pihak saudi arab sangat bijak mempermsudah sarana ibadah haji hanya di indonesia yang di jadikan lahan bisnis oleh pemerintah ini bukan politik tapi ini soal ibadah umat islam
8.	bulan april ini saya dan keluarga akan tunai umrah semoga semuanya lancar serta dipermsudahkan segala urusan umrah nanti
9.	ssudah tertulis apabila pekerjaan jabatan tidak sesuai kemampuan tunggu kehancurannya
10.	berhitungnya jugan dibanding dengan biaya umroh yang hari x â juta naik pesawat pulang pergi jadi haji yang hari harus juta emang berhaji naik pesawatnya pp kan sama naik pesawatnya yang berbeda itu biaya makan hotel dan selama di arofah muzdalifah dan mina

c. Normalisasi Kalimat

Normalisasi kalimat bertujuan untuk mengubah kalimat informal menjadi kalimat formal. Proses ini melibatkan penggantian kata-kata slang dan penghapusan huruf yang berulang, misalnya mengubah "adaaa" menjadi "ada". Untuk mengubah kalimat gaul menjadi kalimat normal, normalisasi kalimat telah dilakukan pada data dokumen, seperti yang ditunjukkan pada Tabel 5.

Tabel 6. Proses Normalisasi Kalimat

No	Kalimat
1.	salam mabrur semoga ada penurunan biar jamaah yang berangkat tahun ini tidak kaget bisa melunasi dan berangkat dengan lega sadarlah banyak yang ekonomi pas pasan
2.	semestinya negara ini bersukur atas kesadarannya rakyatnya patuh dan taat kepada allah penciptanya bila semua manusia patuh terhadap aturan penciptanya dalam alquran dijanjikan negeri ini subur dan makmur serta tidak ada perpecahan di muka bumi ini mestinya negara mendukung dan menfasilitasi dan meringankan beban jamaah haji tidak mencari keuntungan apa lagi memberatkan dengan biaya biaya yang mesti tidak perlu
3.	harusnya kalau mau dinaikan diberlakukannya bagi yang baru daftar jangan yang ssudah daftar apalagi yang ssudah mau berangkat
4.	sebaiknya dpr jangan hanya membahas masalah kenaikan ongkos haji tahun 2023 saja tapi bahas sekaligus untuk ongkos haji tahun berikutnya agar hal hal semacam ini tidak terulang lagi karena selalu membuat kegaduhan keresahan dan memberatkan calon jemaah haji
5.	kecerdasan menteri agama membawa kemudaratan bagi umat muslim di indonesia
6.	yang menaikkan biaya haji semoga allah swt mengazabnya sangat pedih
7.	sebaiknya utk mencapai biaya haji yang ideal itu jangan sekaligus di bebankan tahun sekarang harus bertahap dan berencana dan lamu kunjungan pun bisa dikurangi jangan sampai hari dan tidak kalah pentingnya perlu rencana kenaikan setoran awal yang asalnya rp 25 juta di naikkan misalnya jadi rp 35 juta tapi ini juga harus disosialisasikan jauh jauh
8.	pihak saudi arab sangat bijak mempermsudah sarana ibadah haji hanya di indonesia yang di jadikan lahan bisnis oleh pemerintah ini bukan politik tapi ini soal ibadah umat islam
9.	sebaiknya pengelolaan haji ini diaudit secepatnya sudah kelihatan kezaliman dan mudharatnya dpr bertindaklah sebagai wakil rakyat yang sebenarnya
10.	bulan april ini saya dan keluarga akan tunai umrah semoga

	semuanya lancar serta dipermsudahkan segala urusan umrah nanti

d. *Tokenizing*

Pada langkah *tokenizing* ini, dilakukan pemisahan *string* masukan menjadi potongan-potongan berdasarkan kata-kata yang membentuknya. Intinya, proses ini adalah membagi setiap kata yang membentuk sebuah dokumen. Pada Tabel 6 dapat dilihat kalimat dalam dokumen telah di pecah menjadi kata-kata kemudian menganalisa kumpulan data dengan memisahkan kata tersebut.

Tabel 7. Proses *Tokenizing*

No	Kalimat
1.	['mabrur', 'semoga', 'penurunan', 'biar', 'jamaah', 'berangkat', 'tahun', 'kaget', 'melunasi', 'berangkat', 'lega', 'sadarlah', 'ekonomi', 'pas', 'pasan']
2.	['semestinya', 'negara', 'bersukur', 'kesadarannya', 'rakyatnya', 'patuh', 'taat', 'allah', 'penciptanya', 'manusia', 'patuh', 'aturan', 'penciptanya', 'alquran', 'dijanjikan', 'negeri', 'subur', 'makmur', 'perpecahan', 'muka', 'bumi', 'mestinya', 'negara', 'mendukung', 'menfasilitasi', 'meringankan', 'beban', 'jamaah', 'haji', 'mencari', 'keuntungan', 'memberatkan', 'biaya', 'biaya', 'mesti']
3.	['dinaikan', 'diberlakukannya', 'daftar', 'ssudah', 'daftar', 'ssudah', 'berangkat']
4.	['dpr', 'membahas', 'kenaikan', 'ongkos', 'haji', 'tahun', 'bahas', 'ongkos', 'haji', 'tahun', 'terulang', 'kegaduhan', 'keresahan', 'memberatkan', 'calon', 'jemaah', 'haji']
5.	['kecerdasan', 'menteri', 'agama', 'membawa', 'kemudaratan', 'umat', 'muslim', 'indonesia']
6.	['menaikkan', 'biaya', 'haji', 'semoga', 'allah', 'swt', 'mengazabnya', 'pedih']
7.	['saudi', 'arab', 'bijak', 'mempermsudah', 'sarana', 'ibadah', 'haji', 'indonesia', 'jadikan', 'lahan', 'bisnis', 'pemerintah', 'politik', 'ibadah', 'umat', 'islam']
8.	['april', 'keluarga', 'tunai', 'umrah', 'semoga', 'lancar', 'dipermsudahkan', 'urusan', 'umrah']
9.	['ssudah', 'tertulis', 'pekerjaan', 'jabatan', 'sesuai', 'kemampuan', 'tunggu', 'kehancurannya']
10.	['berhitungnya', 'jugan', 'dibanding', 'biaya', 'umroh', 'â', 'juta', 'pesawat', 'pulang', 'pergi', 'haji', 'juta', 'emang', 'berhaji', 'pesawatnya', 'pp', 'pesawatnya', 'berbeda', 'biaya', 'makan', 'hotel', 'arofah', 'muzdalifah', 'mina']

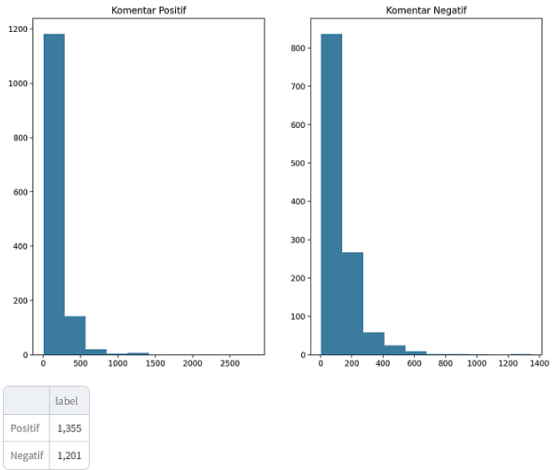
e. *Labelling*

Setelah *preproccessing*, tahapan selanjutnya yaitu pelabelan data yang berupa label positif atau negatif. Dalam melakukan pelebelan data, peneliti menggunakan kamus InSet Lexicon dari penelitian sebelumnya oleh Koto dan Rahmaningtyas [29]. Penelitian ini memiliki nilai akurasi sebesar 65,78 persen. Kamus InSet Lexicon terdiri dari 3.609 kata positif dan 6.609 kata negatif, dan juga memiliki bobot nilai untuk setiap kata. Pada kamus *lexicon* ini, setiap kata dalam bahasa Indonesia memiliki nilai atau *polarity* dengan skor antara -5 dan +5. Contoh hasil pelabelan dapat dilihat di Tabel 7.

Tabel 8. Proses Pelabelan

No	Kalimat	Label
1.	['mabrur', 'semoga', 'penurunan', 'biar', 'jamaah', 'berangkat', 'tahun', 'kaget', 'melunasi', 'berangkat', 'lega', 'sadarlah', 'ekonomi', 'pas', 'pasan']	Positif
2.	['semestinya', 'negara', 'bersukur', 'kesadarannya', 'rakyatnya', 'patuh', 'taat', 'allah', 'penciptanya', 'manusia', 'patuh', 'aturan', 'penciptanya', 'alquran', 'dijanjikan', 'negeri', 'subur', 'makmur', 'perpecahan', 'muka', 'bumi', 'mestinya', 'negara', 'mendukung', 'menfasilitasi', 'meringankan', 'beban', 'jamaah', 'haji', 'mencari', 'keuntungan', 'memberatkan', 'biaya', 'biaya', 'mesti']	
3.	['dinaikan', 'diberlakukannya', 'daftar', 'ssudah', 'daftar', 'ssudah', 'berangkat']	Positif
4.	['dpr', 'membahas', 'kenaikan', 'ongkos', 'haji', 'tahun', 'bahas', 'ongkos', 'haji', 'tahun', 'terulang', 'kegaduhan', 'keresahan', 'memberatkan', 'calon', 'jemaah', 'haji']	Negatif
5.	['kecerdasan', 'menteri', 'agama', 'membawa', 'kemudaratan', 'umat', 'muslim', 'indonesia']	Negatif
6.	['menaikkan', 'biaya', 'haji', 'semoga', 'allah', 'swt', 'mengazabnya', 'pedih']	Negatif
7.	['saudi', 'arab', 'bijak', 'mempermsudah', 'sarana', 'ibadah', 'haji', 'indonesia', 'jadikan', 'lahan', 'bisnis', 'pemerintah', 'politik', 'ibadah', 'umat', 'islam']	Positif
8.	['april', 'keluarga', 'tunai', 'umrah', 'semoga', 'lancar', 'dipermsudahkan', 'urusan', 'umrah']	Positif
9.	['ssudah', 'tertulis', 'pekerjaan', 'jabatan', 'sesuai', 'kemampuan', 'tunggu', 'kehancurannya']	Positif
10.	['berhitungnya', 'jugan', 'dibanding', 'biaya', 'umroh', 'â', 'juta', 'pesawat', 'pulang', 'pergi', 'haji', 'juta', 'emang', 'berhaji', 'pesawatnya', 'pp', 'pesawatnya', 'berbeda', 'biaya', 'makan', 'hotel', 'arofah', 'muzdalifah', 'mina']	Negatif

Pada Gambar 5 merupakan hasil pelabelan data berdasarkan proses *tokenizing*.

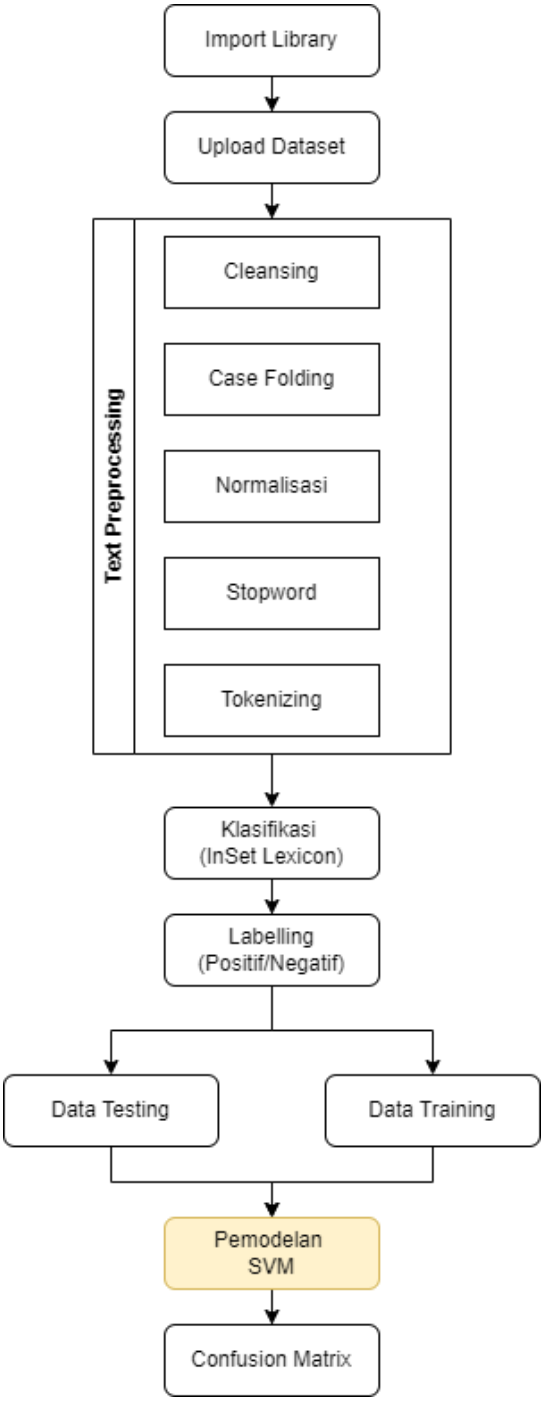


Gambar 5. Pelabelan *Lexicon Based*

3.2 Pemodelan

Penelitian ini melakukan pemodelan menggunakan bahasa pemograman *phyton*. Peneliti melakukan 2 pemodelan, yaitu *Support Vector Machine* dan *Support Vector Machine+Information Gain*.

3.2.1 Support Vector Machine



Gambar 6. Desain Pemodelan SVM

Berikut ini merupakan penjelasan merupakan penjelasan desain pemodelan SVM pada Gambar 6.

1. *Import library*

   Import library yang digunakan pada penelitian ini kurang lebih sebanyak 7 *library*, diantaranya *library* streamlit, numpy, pandas, nltk, SVC, dan lain-lain.
2. *Import data*

   Data yang diperoleh (scrapping) melalui komentar youtube terkait kenaikan biaya ibadah haji 2023. Data yang diperoleh sebanyak 2.500 *dataset*.
3. *Text processing*

   Setelah dilakukan penghapusan data duplikat, selanjutnya diproses melalui text preprocessing. *Cleansing*, *stemming*, *case folding*, *stopword*, normalisasi, dan *tokenizing* adalah metode *text preprocessing* yang digunakan dalam penelitian ini.
4. *InSet Lexicon*

   Pada tahap ini, setiap komentar dianalisis satu per satu. menggunakan kamus *InSet Lexicon*. Selanjutnya, polarity score dihitung untuk setiap kalimat. Ini dilakukan dengan cara menjumlahkan bobot total dari kata yang diidentifikasi oleh sistem. Selanjutnya, data komentar dikategorikan berdasarkan jenis sentimen, yaitu positif atau negatif.
5. *Labelling*

   Kalimat komentar dianggap positif jika bobot *polarity score*-nya lebih besar dari 0 dan negatif jika bobot *polarity score*-nya kurang dari 0.
6. *Data Training* dan *Data Testing*

   Dari 2.500 dataset dibagi menjadi dua bagian yaitu data training (80%) dan data testing (20%).
7. SVM

   Pemodelan metode SVM diimplementasi untuk diuji hasil akurasinya. Hasil akurasi yang diperoleh sebesar 87%.
8. *Confusion Matrix*

   Tahap terakhir yaitu mengukur performa klasifikasi menggunakan confusion matrix.

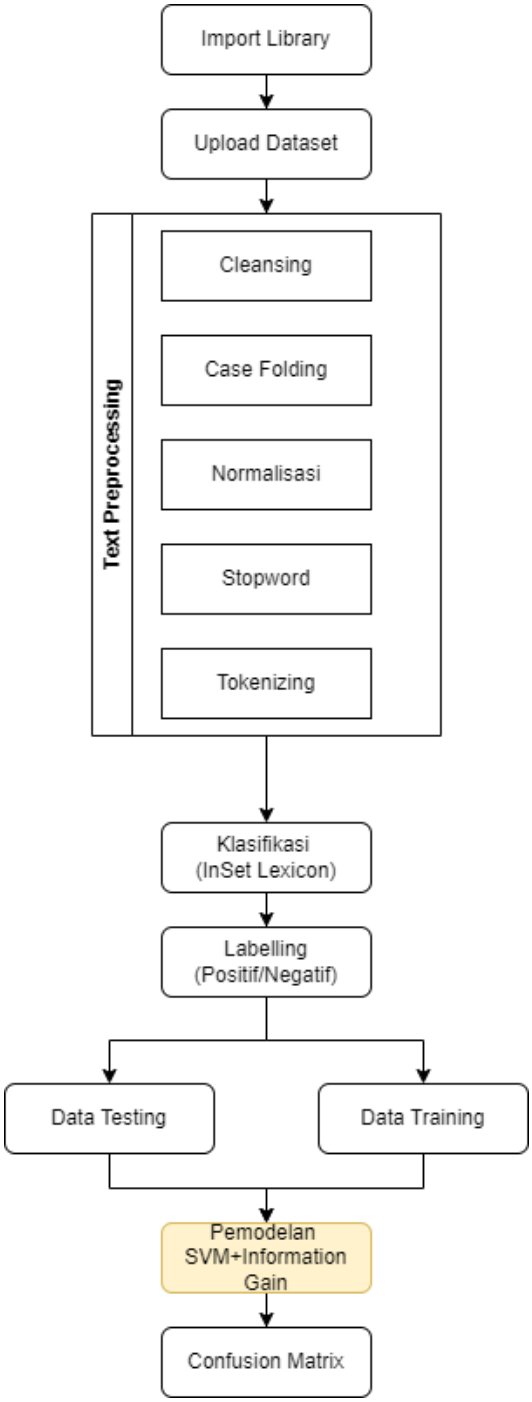
3.2.2 Support Vector Machine + Information Gain

Berikut ini merupakan penjelasan merupakan penjelasan desain pemodelan SVM pada Gambar 7.

1. *Import library*

   Import library yang digunakan pada penelitian ini kurang lebih sebanyak 7 *library*, diantaranya *library streamlit*, *numpy*, *pandas*, nltk, SVC, dan lain-lain.
2. *Import data*

   Data yang diperoleh *(scrapping)* melalui komentar youtube terkait kenaikan biaya ibadah haji 2023. Data yang diperoleh sebanyak 2.500 *dataset*.



Gambar 7. Desain Pemodelan SVM + Information Gain

3. *Text processing*
   Setelah dilakukan penghapusan data duplikat, selanjutnya diproses melalui text preprocessing. *Cleansing*, *stemming*, *case folding*, *stopword*, normalisasi, dan *tokenizing* adalah metode *text preprocessing* yang digunakan dalam penelitian ini.
4. *InSet Lexicon*
   Pada tahap ini, setiap komentar dianalisis satu per satu. menggunakan kamus *InSet Lexicon*. Selanjutnya, polarity score dihitung untuk setiap kalimat. Ini dilakukan dengan cara menjumlahkan bobot total dari kata yang diidentifikasi oleh sistem. Selanjutnya, data komentar dikategorikan berdasarkan jenis sentimen, yaitu positif atau negatif.
5. *Labelling*
   Kalimat komentar dianggap positif jika bobot *polarity score*-nya lebih besar dari 0 dan negatif jika bobot *polarity score*-nya kurang dari 0.
6. *Data Training* dan *Data Testing*
   Dari 2.500 dataset dibagi menjadi dua bagian yaitu data training (80%) dan data testing (20%).
7. SVM+*Information Gain*
   Pemodelan metode SVM+*Information Gain* diimplementasi untuk diuji hasil akurasinya. Hasil akurasi yang diperoleh sebesar 89%.
8. *Confusion Matrix*
   Tahap terakhir yaitu mengukur performa klasifikasi menggunakan *confusion matrix*.

3.3. Evaluasi

Setelah dilakukan pemodelan selanjutnya hasil evaluasi dari kedua model menggunakan *confusion matrix* untuk mengukur performa.

3.3.1 Confusion Matrix SVM

Hasil yang dihasilkan saat menguji metode SVM dengan komposisi data latih dan data uji 80:20 adalah sebagai berikut:

Tabel 9. *Confusion Matrix* Pemodelan SVM

Confusin Matrix		**Predicted Class**	
		Positif	**Negatif**
Observed Class	**Positif**	TP = 195	FN = 44
	Negatif	FP = 23	TN = 250

$$\text{Recall (Negatif)} = \frac{TP}{TP+FP} = \frac{195}{195+23} = \frac{195}{218} = 0.894\ x\ 100\% = 89.4\%$$

Dibulatkan menjadi = 89%

$$\text{Recall (Positif)} = \frac{TN}{TN+FN} = \frac{250}{250+44} = \frac{250}{294} = 0.850\ x\ 100\% = 85.0\%$$

Dibulatkan menjadi = 85%

$$\text{Precision (Negatif)} = \frac{TP}{TP+FN} = \frac{195}{195+44} = \frac{195}{239} = 0.815\ x\ 100\% = 81.5\%$$

Dibulatkan menjadi = 82%

Precision (Positif) = $\frac{TN}{TN+FP} = \frac{250}{250+23} = \frac{250}{273} = 0.915\ x\ 100\% = 91.5\%$

Dibulatkan menjadi = 92%

Akurasi = $\frac{TP+TN}{TP+FP+TN+FN} = \frac{195+250}{195+23+250+44} = \frac{445}{512} = 0.869\ x\ 100\% = 86.9\%$

Dibulatkan menjadi = 87%

3.3.2 Confusion Matrix SVM+*Information Gain*

Hasil yang dihasilkan dari pengujian metode SVM+*Information Gain* dengan komposisi data latih dan data uji 80:20 adalah sebagai berikut:

Tabel 10. Confusion Matrix Pemodelan SVM+Information Gain

Confusin Matrix		Predicted Class	
		Positif	**Negatif**
Observed Class	**Positif**	TP = 196	FN = 36
	Negatif	FP = 22	TN = 258

Recall (Negatif) = $\frac{TP}{TP+FP} = \frac{196}{196+22} = \frac{196}{218} = 0.899\ x\ 100\% = 89.9\%$

Dibulatkan menjadi = 90%

Recall (Positif) = $\frac{TN}{TN+FN} = \frac{258}{258+36} = \frac{258}{294} = 0.877\ x\ 100\% = 87.7\%$

Dibulatkan menjadi = 88%

Precision (Negatif) = $\frac{TP}{TP+FN} = \frac{196}{196+36} = \frac{196}{232} = 0.844\ x\ 100\% = 84.4\%$

Dibulatkan menjadi = 84%

Precision (Positif) = $\frac{TN}{TN+FP} = \frac{258}{258+22} = \frac{258}{273} = 0.945\ x\ 100\% = 94.5\%$

Dibulatkan menjadi = 95%

Akurasi = $\frac{TP+TN}{TP+FP+TN+FN} = \frac{196+258}{196+22+258+36} = \frac{454}{512} = 0.886\ x\ 100\% = 88.6\%$

Dibulatkan menjadi = 89%

4. DISKUSI

4.1. Analisis Hasil Pengujian

Berdasarkan kedua pemodelan, selanjutnya dibandingkan dan diverifikasi untuk menguji nilai akurasinya, terlihat bahwa algoritma SVM yang dikombinasikan dengan *information gain* memberikan tingkat akurasi yang lebih tinggi daripada algoritma SVM tanpa dikombinasikan dengan *information gain*.

Tabel 11. Hasil Perbandingan Akurasi dari Dua Model

No	Metode	Akurasi	Recall (negatif)	Recall (positif)	Precision (negatif)	Precision (positif)
1	SVM	87%	89%	85%	82%	92%
2	SVM+ IG	89%	90%	88%	84%	95%

Dari hasil pengujian pada Tabel 9, didapat hasil bahwa algoritma SVM yang dikombinasikan dengan *information gain* memiliki nilai akurasi tertinggi sebesar 89%, nilai recall negatif sebesar 90%, nilai recall positif sebesar 88%, nilai *precision* negatif sebesar 84%, nilai precision positif sebesar 95%. Jika hasil dari studi ini dibandingkan dengan penelitian sebelumnya yang dilakukan oleh Larasati, dkk (2019), SVM yang dikombinasikan dengan fitur Chi Square dan TFIDF menghasilkan akurasi sebesar 80.2%.

5. KESIMPULAN

Algoritma SVM yang dikombinasikan dengan fitur *information gain* terbukti memiliki kinerja yang lebih baik dengan akurasi mencapai 89%, dibandingkan dengan SVM tanpa dikombinasikan fitur *information gain* yang mencapai akurasi sebesar 87%.

DAFTAR PUSTAKA

[1] R. Asrianto and M. Herwinanda, "Analisis sentimen kenaikan harga kebutuhan pokok dimedia sosial youtube menggunakan algoritma support vector machine," *J. CoSciTech (Computer Sci. Inf. Technol.*, vol. 3, no. 3, pp. 431–440, Dec. 2022, doi: 10.37859/coscitech.v3i3.4368.

[2] katadata.co.id, "Facebook hingga Twitter, Ini Deretan Media Sosial Terpopuler Dunia di Awal 2023," 2023. Accessed: Feb. 06, 2023. [Online]. Available: https://databoks.katadata.co.id/datapublish/2023/02/06/facebook-hingga-twitter-ini-deretan-media-sosial-terpopuler-dunia-di-awal-2023

[3] Kemenag, "Keputusan Menteri Agama No 352 Tahun 2023 tentang Biaya Perjalanan Ibadah Haji (Bipih) Reguler 1444 H dan Penggunaan Nilai Manfaat," 2023. https://kemenag.go.id/informasi/keputusan-

menteri-agama-no-352-tahun-2023-tentang-biaya-perjalanan-ibadah-haji--bipih--reguler-1444-h-dan-penggunaan-nilai-manfaat

[4] D. M. Y. Sinurat, D. E. Ratnawati, and D. W. Brata, "Analisis Sentimen Terhadap Kenaikan Cukai Rokok pada Media Sosial Twitter menggunakan Algoritma Naïve Bayes Classifier," 2023. [Online]. Available: http://j-ptiik.ub.ac.id

[5] F. V. Sari and A. Wibowo, "Analisis Sentimen Pelanggan Toko Online JD.ID Menggunakan Metode Naive Bayes Classifier Berbasis Konversi Ikon Emosi," *J. SIMETRIS*, vol. 10, no. 2, 2019.

[6] R. K. Dey, D. Sarddar, I. Sarkar, R. Bose, and S. Roy, "Techniques Involving Social Media And Online Platforms," *Int. J. Sci. Technol. Res.*, vol. 9, no. 05, 2020.

[7] A. Giachanou and F. Crestani, "Like it or not: A survey of Twitter Sentiment Analysis Methods," *ACM Computing Surveys*, vol. 49, no. 2. Association for Computing Machinery, Jun. 01, 2016. doi: 10.1145/2938640.

[8] D. Juhaeni and A. Wibowo, "Penerapan Metode Naïve Bayes Untuk Wacana Kenaikan Harga Tiket Candi Borobudur Pada Twitter," 2022.

[9] S. Nurhaliza, Y. Yusra, and M. Fikry, "Klasifikasi Sentimen Masyarakat di Twitter Terhadap Kenaikan Harga BBM dengan Metode Support Vector Machine," *J. Sist. Komput. dan Inform.*, vol. 4, no. 4, p. 586, Jul. 2023, doi: 10.30865/json.v4i4.6322.

[10] R. Savira, A. Solichin, and M. Syafrullah, "Analisis Sentimen Pada Twitter Terhadap KenaikanBBM 2022 Dengan Lexicon dan Support Vector Machine," Jakarta, 2023.

[11] A. S. H. Basari, B. Hussin, I. G. P. Ananta, and J. Zeniarja, "Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization," in *Procedia Engineering*, Elsevier Ltd, 2013, pp. 453–462. doi: 10.1016/j.proeng.2013.02.059.

[12] L. B. Ilmawan and M. A. Mude, "Perbandingan Metode Klasifikasi Support Vector Machine dan Naïve Bayes untuk Analisis Sentimen pada Ulasan Tekstual di Google Play Store," *Ilk. J. Ilm.*, vol. 12, no. 2, pp. 154–161, 2020, doi: 10.33096/ilkom.v12i2.597.154-161.

[13] W. Athira Luqyana, I. Cholissodin, and R. S. Perdana, "Analisis Sentimen Cyberbullying pada Komentar Instagram dengan Metode Klasifikasi Support Vector Machine," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 11, pp. 4704–4713, 2018, [Online]. Available: http://j-ptiik.ub.ac.id

[14] D. Wang and Y. Zhao, "Using News to Predict Investor Sentiment: Based on SVM Model," *Procedia Comput. Sci.*, vol. 174, no. 2019, pp. 191–199, 2020, doi: 10.1016/j.procs.2020.06.074.

[15] F. D. Ananda and Y. Pristyanto, "Analisis Sentimen Pengguna Twitter Terhadap Layanan Internet Provider Menggunakan Algoritma Support Vector Machine," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 20, no. 2, pp. 407–416, May 2021, doi: 10.30812/matrik.v20i2.1130.

[16] P. Arsi, R. Wahyudi, and R. Waluyo, "Optimasi SVM Berbasis PSO pada Analisis Sentimen Wacana Pindah Ibu Kota Indonesia," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 2, pp. 231–237, 2021, doi: 10.29207/resti.v5i2.2698.

[17] L. P. Hung, R. Alfred, and M. H. A. Hijazi, "A review on feature selection methods for sentiment analysis," *Advanced Science Letters*, vol. 21, no. 10. American Scientific Publishers, pp. 2952–2956, Oct. 01, 2015. doi: 10.1166/asl.2015.6475.

[18] V. Chandani and R. S. Wahono, "Komparasi Algoritma Klasifikasi Machine Learning Dan Feature Selection pada Analisis Sentimen Review Film," *J. Intell. Syst.*, vol. 1, no. 1, 2015, [Online]. Available: http://journal.ilmukomputer.org

[19] A. B. P. Negara, H. Muhardi, and I. M. Putri, "Analisis Sentimen Maskapai Penerbangan Menggunakan Metode Naive Bayes dan Seleksi Fitur Information Gain," vol. 7, no. 3, pp. 599–606, 2020, doi: 10.25126/jtiik.202071947.

[20] M. Arhami and M. Nasir, *Data Mining Algoritma dan Implementasi*, 1st ed. Yogyakarta: Andi, 2020.

[21] Ratih Puspitasari, Y. Findawati, and M. A. Rosid, "SENTIMENT ANALYSIS OF POST-COVID-19 INFLATION BASED ON TWITTER USING THE K-NEAREST NEIGHBOR AND SUPPORT VECTOR MACHINE CLASSIFICATION METHODS," *J. Tek. Inform.*, vol. 4, no. 4, pp. 669–679, Aug. 2023, doi: 10.52436/1.jutif.2023.4.4.801.

[22] F. A. Sianturi, P. M. Hasugian, A. Simangunsong, and B. Nadeak, *Data Mining : Teori dan Aplikasi Weka*, 1st ed. CV. Rudang Mayang, 2019.

[23] D. A. Pisner and D. M. Schnyer, *Support vector machine*. Elsevier Inc., 2020. doi: 10.1016/B978-0-12-815739-8.00006-7.

[24] H. Mubarok, I. Ernawati, and N. Chamidah,

"Optimasi Algoritma Support Vector Machine Menggunakan Seleksi Fitur Particle Swarm Optimization Pada Analisis Sentimen Terhadap Kebijakan PPKM," pp. 1–54, 2022.

[25] I. S. Ahmad, A. A. Bakar, and M. R. Yaakub, "A Review Of Feature Selection in Sentiment Analysis Using Information Gain and Domain Specific Ontology," *Int. J. Adv. Comput. Res.*, vol. 9, no. 44, pp. 283–292, Sep. 2019, doi: 10.19101/ijacr.pid90.

[26] G. Wu and J. Xu, "Optimized Approach of Feature Selection based on Information Gain," in *Proceedings - 2015 International Conference on Computer Science and Mechanical Automation, CSMA 2015*, Institute of Electrical and Electronics Engineers Inc., Jan. 2016, pp. 157–161. doi: 10.1109/CSMA.2015.38.

[27] D. Musfiroh, U. Khaira, P. E. P. Utomo, and T. Suratno, "Analisis Sentimen terhadap Perkuliahan Daring di Indonesia dari Twitter Dataset Menggunakan InSet Lexicon," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 1, no. 1, pp. 24–33, 2021, doi: 10.57152/malcom.v1i1.20.

[28] I. K. A. B. Artana, G. A. Pradnyana, and I. G. M. Darmawiguna, "ANALISIS SENTIMEN TWITTER UNTUK MENILAI KESIAPAN PEMBELAJARAN TATAP MUKA TERBATAS DENGAN INSET LEXICON DAN LEVENSHTEIN DISTANCE," *J. Pendidik. Teknol. dan Kejuru.*, vol. 20, no. 2, 2023.

[29] F. Koto and G. Y. Rahmaningtyas, "InSet Lexicon: Evaluation of a Word List for Indonesian Sentiment Analysis in Microblogs," pp. 391–394, 2017, doi: 10.1109/IALP.2017.8300625.

[30] S. Visa, A. Inoue, and A. Ralescu, "Proceedings of the Twenty--second Midwest Artificial Intelligence and Cognitive Science Conference," 2011.

[31] E. Prasetyowati, *Data Mining : Pengelompokan Data untuk Informasi dan Evaluasi*. Pamekasan: Duta Media Publishing, 2017.

[32] E. A. Novia, W. I. Rahayu, and C. Prianto, *Sistem Perbandingan Algoritma K-Means Dan Naive Bayes Untuk Memprediksi Prioritas Pembayaran Tagihan Rumah Sakit Berdasarkan Tingkat Kepentingan*. Bandung: Kreatif Industri Nusantara, 2020.