

LAPORAN FINAL PROJECT

**Analisis Sentimen Tempat Wisata di Daerah Istimewa  
Yogyakarta Menggunakan Metode Support Vector Machine**

Disusun untuk memenuhi Ujian Akhir Semester

Mata Kuliah : Data Mining

Dosen Pengampu : Kusnawi, S.Kom, M.Eng



Oleh :

**Aqsal Harris Pratama                      19.11.3201**

**Ihdi Ulhaq Abdul Al Aslam                19.11.3239**

**PROGRAM STUDI S1 INFORMATIKA  
UNIVERSITAS AMIKOM YOGYAKARTA  
JANUARI 2022**

## **BAB 1. PENDAHULUAN**

### **1.1 Latar Belakang Masalah**

Saat ini teknologi *artificial intelligence* sudah diterapkan dalam berbagai sektor kehidupan, salah satunya adalah sektor pariwisata. Pemanfaatan *artificial intelligence* pada sektor wisata bertujuan untuk membantu pengelola tempat wisata dalam membuat keputusan dan mengambil suatu tindakan berdasarkan hasil yang diolah oleh teknologi tersebut. Salah satu hal yang dapat diolah oleh teknologi *artificial intelligence* adalah ulasan pengunjung pada situs penyedia informasi tempat seperti Google Maps.

Google Maps dapat mencari dan menampilkan informasi mengenai suatu tempat. Informasi yang ditampilkan didapatkan dari masukan-masukan oleh pengunjung dan pemilik tempat tersebut. Salah satu informasi yang dapat ditampilkan adalah ulasan pengunjung. Berdasarkan ulasan tersebut calon pengunjung dapat mengetahui bagaimana pendapat orang lain mengenai tempat tersebut.

Menurut [1], 78% pengunjung lebih mempercayai ulasan online daripada rekomendasi seseorang. Sehingga, dapat dikatakan bahwa ulasan pengunjung pada berbagai platform merupakan hal yang cukup penting untuk dipertimbangkan bagi pihak pengelola tempat wisata. Namun, semakin banyaknya ulasan yang ada akan mempersulit pengelola untuk mengetahui pokok ulasan dari seluruh pengunjung. Oleh karena itulah perlu untuk diterapkannya salah satu bidang dari teknologi *artificial intelligence* yang disebut sebagai *Natural Language Processing* dengan aplikasi berupa analisis sentimen.

Analisis sentimen dapat digunakan memperoleh opini dari pengguna sehingga akan membantu pihak pengelola dalam memperoleh inti dari berbagai ulasan secara efisien. Dalam penelitian ini, analisis sentimen diterapkan untuk mengetahui jenis ulasan pengunjung apakah positif atau negatif dan aspek yang dibahas pada ulasan tersebut seperti kebersihan, daya tarik, fasilitas, dan pelayanan.

Penelitian [2] menunjukkan bahwa terdapat beberapa metode penyelesaian masalah analisis sentimen di bawah *supervised learning* seperti *Support Vector*

*Machine* (SVM), *Naive Bayes Classifier* (NBC), dan *Maximum Entropy* (ME). SVM bekerja dengan mengklasifikasikan objek berdasarkan *feature* yang ada dan menentukan posisinya terhadap *hyperlane*. Sedangkan NBC dan ME bekerja dengan menghitung probabilitas suatu kelas terhadap seluruh dokumen.

Pada penelitian ini algoritma yang dipilih adalah SVM karena cocok untuk diterapkan dengan pembobotan TF-IDF yang memberi bobot terhadap setiap kata pada suatu kalimat yang menjadi *feature* pada SVM. Selain itu, menurut [3] SVM memiliki akurasi yang relatif tinggi.

Dengan penelitian ini diharapkan dapat menjadi suatu model dalam analisis sentimen tempat wisata bagi pengelola tempat wisata dan dikembangkan dalam penelitian lain.

## **1.2 Rumusan Masalah**

Berdasarkan uraian latar belakang yang ada, maka rumusan masalah yang menjadi fokus pada penelitian ini adalah sebagai berikut.

1. Apakah algoritma SVM dapat diterapkan dalam analisis sentimen berdasarkan ulasan pada Google Maps Review?
2. Berapa akurasi yang dihasilkan SVM dalam analisis sentimen berdasarkan ulasan pada Google Maps Review?

## **1.3 Batasan Masalah**

Beberapa batasan masalah dalam penelitian ini adalah sebagai berikut.

1. Data yang digunakan adalah data dari Google Maps Review.
2. Data yang digunakan terbatas pada data ulasan tempat wisata di Daerah Istimewa Yogyakarta pada Google Maps Review.

## **1.4 Maksud dan Tujuan**

Maksud dan tujuan dari penelitian ini adalah sebagai berikut.

1. Mengetahui performa algoritma SVM dalam analisis sentimen berdasarkan ulasan pada Google Maps Review.
2. Mengetahui akurasi yang dihasilkan SVM dalam analisis sentimen berdasarkan ulasan pada Google Maps Review.

## **1.5 Manfaat Penelitian**

Hasil dari penelitian ini diharapkan menjadi manfaat bagi :

1. Pembaca untuk mengetahui penerapan dan manfaat analisis sentimen pada ulasan tempat wisata.
2. Pengelola tempat wisata dalam membuat keputusan dan kebijakan berdasarkan ulasan pengunjung.
3. Pengembang aplikasi mengenai tempat wisata sebagai suatu model dalam pembuatan aplikasi.

## BAB 2. DASAR TEORI

Pada era saat ini data mining bermain peran yang sangat peting dalam perkembangan industry terutama pada area text mining [4]. Data mining adalah sebuah proses untuk mengekstrak sebuah informasi dari data data untuk mengetahui pola pada sebuah kumpulan data yang besar [4]. Dengan ditemukanya informasi dan pola dalam kumpulan data tersebut maka sebuah trend, pattern, relationship dan route akan terbantu dan kemudian diolah untuk memenuhi kebutuhan perusahaan.

Seiring perkembangan zaman dengan makin maraknya peneliti dibidang artificial intelegence dan data science teknologi text mining menjadi perhatian tersendiri dimata para peneliti. Text mining memfokuskan kaidah text analisis dan computer science untuk menghandle masalah yang kompleks [5]. Text mining dapat menemukan dan menerima sebuah informasi dalam sebuah kumpulan data atau text corpus. Text mining mengkombinasikan artificial intelligence, information retrieval dan data mining dalam memahami *analytical processing systems*.

Perkembangan e-commerce menuntut semua perusahaan harus mengembangkan usahanya. Dengan menggunakan sentiment analis perusahaan akan mendapatkan kemampuan untuk melihat target pasar mereka dengan ini sentiment analis akan memperkuat product dan service perusahaan terkait. Sentiment analisis merupakan sebuah bidang yang berkaitan dengan language processing dan dengan adanya sentiment analis ini pihak perusahaan maupun industri dapat merancang dan mengidentifikasi kepuasan pelanggan. Sebuah aspek dalam sentiment analisis merujuk pada identifikasi sebuah ekspresi untuk mengekstrak *fine-grained information* [6].

Penelitian [7] menerapkan gabungan teknik *Lexicon* dan *Machine Learning* dalam penyelesaian masalah analisis sentimen. Teknik *Lexicon* diterapkan pada bagian pelabelan dokumen dan *Machine Learning* dalam proses klasifikasi sentimen. Dengan menerapkan kedua teknik tersebut akan memudahkan proses analisis sentimen terhadap suatu kasus.

Penelitian [8] menggunakan *Word2Vec* dalam proses ekstraksi fitur atau representasi vektor dari tiap kata. Hasilnya menunjukkan bahwa ekstraksi fitur menggunakan *Word2Vec* memiliki akurasi terendah dibandingkan dengan model lainnya, yang mana *Word2Vec* menghasilkan akurasi 70% sedangkan model lainnya lebih dari 80%. Hal tersebut dikarenakan *Word2Vec* memerlukan lebih banyak data agar mencapai hasil yang lebih optimal.

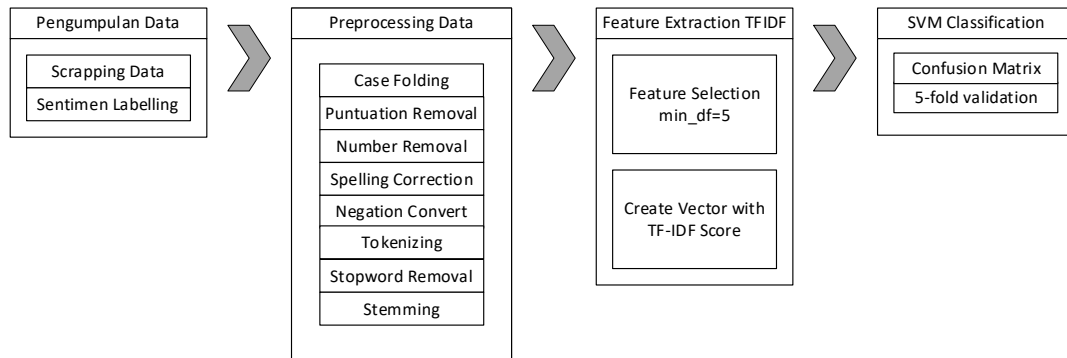
Proses *stopword removal* dan *stemming* merupakan salah satu proses dalam *text preprocessing* sebelum ekstraksi fitur dalam penyelesaian masalah analisis sentimen. Penelitian [9] membandingkan pendekatan yang berbeda dalam proses tersebut, seperti *stop – stem*, *no stop – stem*, *stop – no stem*, dan *no stop – no stem*. Hasilnya menunjukkan bahwa proses *stopword removal* dan *stemming* tidak berpengaruh banyak terhadap akurasi yang dihasilkan setelah proses klasifikasi dilakukan.

Masalah analisis sentimen dapat diselesaikan menggunakan beberapa algoritma. Penelitian [10] menggunakan berbagai algoritma dan masing-masing algoritma tersebut menghasilkan akurasi yang relatif tinggi. SVM menghasilkan akurasi tertinggi yaitu sebesar 97.72% dibandingkan dengan algoritma lainnya.

## BAB 3. TEKNIK DATA MINING

### 3.1 Alur Penelitian

Penelitian ini terdiri atas empat tahapan, dimulai dari tahap pengumpulan data, *text preprocessing*, ekstraksi fitur, dan klasifikasi menggunakan SVM. Adapun alur penelitian ditunjukkan sebagai berikut pada Gambar 1.



Gambar 1. Diagram alur penelitian

#### 3.3.1 Pengumpulan Data

Penelitian ini menggunakan data ulasan pada berbagai tempat wisata di DIY seperti Candi Prambanan, Gembira Loka, Hartono Mall, Jogja Bay, Pantai Depok, Pantai Parangtritis, dan Tebing Breksi dari Google Maps Review. Data dari tempat-tempat tersebut di-*scrapping* menggunakan Apify kemudian digabungkan.

Keseluruhan data berjumlah 2434 data yang kemudian diberi label sentimen sebagai ulasan positif dan negatif berdasarkan bintang masing-masing ulasan yang kemudian dicek ulang secara manual. Data yang sudah diberi label menunjukkan ulasan positif berjumlah 1588 data dan ulasan negatif berjumlah 846. Jumlah data tersebut menandakan bahwa data yang digunakan *imbalance* atau tidak seimbang.

Tabel 1. Contoh Dataset

text	value
------	-------

Banyak resto seafood/boga bahari, tempat bagus, dan cocok untuk menikmati senja bareng keluarga atau kekasih ðŸ‘°ðŸ‘°ðŸ‘°~,	POSITIF
pantai kotor	NEGATIF
Enak buat makan seafood	POSITIF
Tempat nya bagus, hanya saja untuk saat ini masih banyak wahana yang masih tutup. Karna efek pandemi.	POSITIF
Aku benci masa 1 orang bayarnya mahal bankrut dah!!	NEGATIF

### 3.3.2 Text Preprocessing

*Preprocessing* dilakukan dengan tujuan agar data menjadi bersih dan siap untuk diproses pada tahap selanjutnya [11]. Tahapan pada *Text Preprocessing* yang akan dilakukan pada penelitian ini meliputi *casefolding* (mengubah huruf kapital menjadi huruf kecil), *punctuation removal* (menghilangkan tanda baca dan karakter selain alfabet), *number removal* (menghilangkan angka), *tokenizing* (memisahkan setiap kata pada kalimat menjadi token), *stopword removal* (menghilangkan kata yang tidak dibutuhkan), dan *stemming* (mengubah kata berimbuhan menjadi kata dasar).

Tabel 2. Contoh *Text Preprocessing*

text	value
Teks Awal	Aku benci masa 1 orang bayarnya mahal bankrut dah!!
Case Folding	aku benci masa 1 orang bayarnya mahal bankrut dah!!
Punctuation Removal	aku benci masa 1 orang bayarnya mahal bankrut dah
Number Removal	aku benci masa orang bayarnya mahal bankrut dah
Tokenizing	[aku, benci, masa, orang, bayarnya, mahal, bankrut, dah]
Stopword Removal	[benci, orang, bayarnya, mahal, bankrut]
Stemming	[benci, orang, bayar, mahal, bankrut]



### 3.3.3 Ekstraksi Fitur

Proses ini akan mengubah setiap kata berbentuk token yang sudah melalui tahap *preprocessing* menjadi vektor yang akan merepresentasikan kata yang ada.

*Term Frequency* (TF) merupakan frekuensi kata yang muncul pada sebuah dokumen. *Inverse Document Frequency* (IDF) mengukur seberapa pentingnya kata pada sebuah dokumen [12]. Adapun persamaan dalam penghitungan TF-IDF :

Bobot kata  $i$  pada dokumen  $j$  :

$$W_{i,j} = tf_{i,j} \times \log \left( \frac{N}{df_i} \right)$$

$tf_{i,j}$  = jumlah kata  $i$  muncul dalam dokumen  $j$

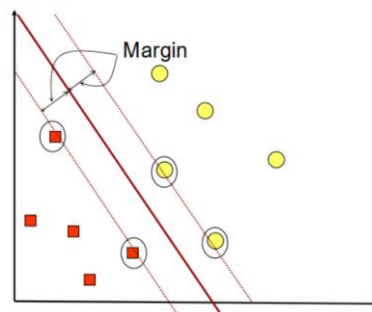
$df_i$  = jumlah dokumen yang mengandung  $i$

$N$  = total seluruh dokumen

### 3.3.4 Klasifikasi SVM

#### 3.3.4.1 Support Vector Machine

SVM merupakan algoritma *machine learning* dengan tipe *supervised* berbasis vektor. SVM mengklasifikasikan data dengan membaginya menjadi dua kelas melalui *hyperplane* [10]. *Hyperplane* berada di antara dua kelas dengan jarak  $d$  pada titik terdekat setiap kelas. Jarak  $d$  ini disebut dengan *margin* dan titik yang berada pada *margin* ini disebut sebagai *support vector*. Tujuan dari SVM adalah menentukan *hyperplane* terbaik yang dapat memberikan jarak terjauh dari suatu titik [13]. Representasi SVM dinyatakan pada gambar berikut [10].



Gambar 2. Representasi SVM

#### 3.3.4.2 Confusion Matrix

*Confusion Matrix* merupakan salah satu metode dalam evaluasi performa suatu model dalam bentuk tabel yang menampilkan hasil prediksi terhadap dua kelas sebagai berikut.

Tabel 3. *Confusion Matrix*

		Kelas Asli	
		Kelas-1	Kelas-2
Kelas Prediksi	Kelas-1	True Positive (TP)	False Negative (FN)
	Kelas-2	False Positive (FP)	True Negative (TN)

Adapun fungsi dari *confusion matrix* ini adalah mengukur tingkat akurasi dari suatu model dengan persamaan sebagai berikut.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

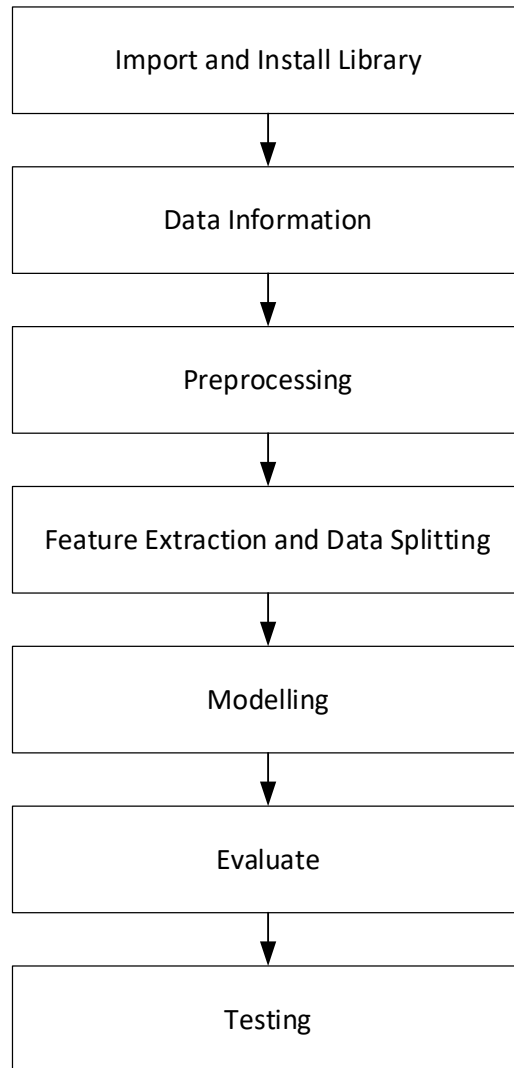
#### 3.3.4.3 K-Fold Cross Validation

Evaluasi yang lebih mendalam dapat dilakukan dengan *k-fold cross-validation*. Metode ini akan membagi dataset secara acak sebanyak “*k*” yang ditentukan secara sama besar [14]. Metode ini menguji stabilitas model yang digunakan terhadap suatu data dan menghasilkan rata-rata akurasi. Pada penelitian ini, nilai “*k*” yang digunakan adalah 10.

## BAB 4. IMPLEMENTASI DAN PEMBAHASAN

### 4.1 Alur Program

Berikut ini adalah alur program pada Google Colab yang telah dibuat.



Gambar 2. Alur Program

#### 4.1.1 Import and Install Library

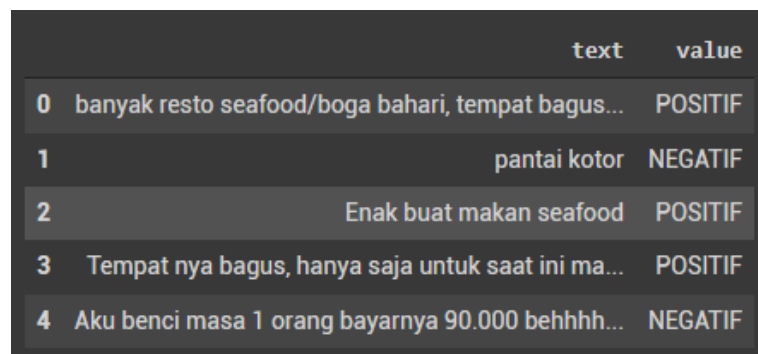
Dalam tahapan ini program akan melakukan importing library dengan keterangan library seperti : Pandas, String, Sklearn, Matplotlib, cross\_val\_score, nltk, re, sastrawi. Penggunaan library memungkinkan proses pengodingan dan penelitian akan menjadi lebih mudah.

#### 4.1.2 Data Information

Pada tahapan ini program akan memproses dataset dengan cara membaca sebuah path atau lokasi yang telah inputkan oleh pemrogram.

##### 4.1.2.1 Dataset Read

Dalam tahapan ini dataset yang telah terbaca oleh program akan di import dan kemudian ditampilkan dalam tampilan



	text	value
0	banyak resto seafood/boga bahari, tempat bagus...	POSITIF
1	pantai kotor	NEGATIF
2	Enak buat makan seafood	POSITIF
3	Tempat nya bagus, hanya saja untuk saat ini ma...	POSITIF
4	Aku benci masa 1 orang bayarnya 90.000 behhhh...	NEGATIF

Gambar 3. Membaca Datase

##### 4.1.2.2 General Information

Dalam tahapan ini program akan menampilkan informasi dataset yang akan diolah. Program akan memberikan informasi tentang jumlah kolom dan baris dalam dataset tersebut. Dalam tahapan ini juga diberikan detail info dari dataset dengan diberikan keterangan index, column, Non-Null Count, Dtype.

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2434 entries, 0 to 2433
Data columns (total 2 columns):
#   Column  Non-Null Count  Dtype
---  -
0    text    2434 non-null    object
1    value    2434 non-null    object
dtypes: object(2)
memory usage: 38.2+ KB

```

Gambar 4. Data Information

Dataset juga akan ditampilkan dengan menggunakan tabel yang berisi jumlah baris, jumlah baris unik, kata teratas dalam dataset, dan frekuensi pada dataset.

	text	value
count	2434	2434
unique	2343	2
top	Bagus	POSITIF
freq	12	1588

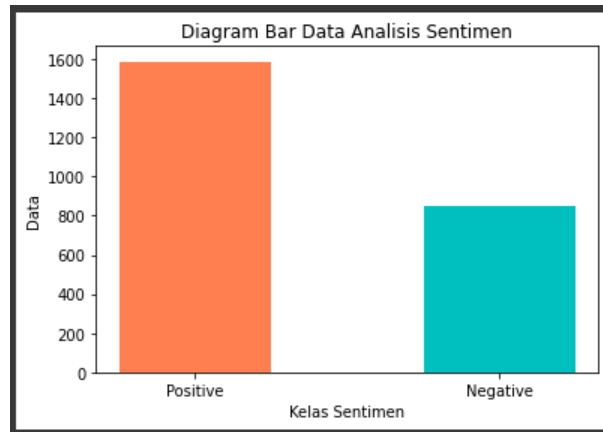
Gambar 5. Data Describe

#### 4.1.2.3 Check Missing Values

Dalam tahapan ini program akan menampilkan data yang bersifat null dalam dataset.

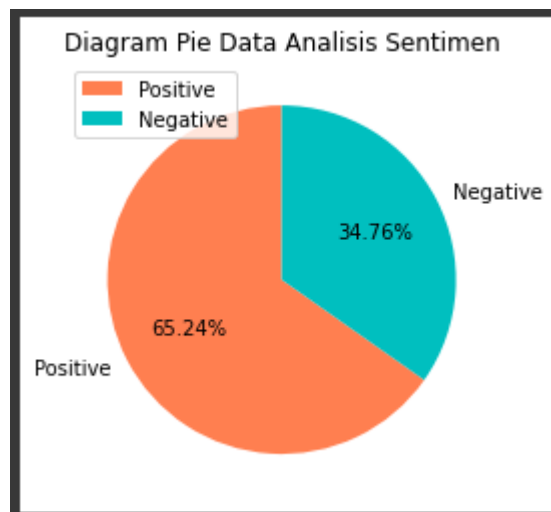
#### 4.1.2.4 Visualize Data

Dataset yang telah dimasukkan kedalam program akan diolah dan dijadikan diagram batang untuk mengukur tingkat polaritas dataset.



Gambar 6. Visualisasi Data Diagram Batang

Program juga akan menampilkan diagram pie dengan menampilkan perbandingan dalam bentuk persen sebagai kondisi polaritas dataset.



Gambar 7. Visualisasi Data Diagram Pie

#### 4.1.3 Preprocessing

Tahap ini adalah proses pembersihan dataset atau pengolahan dataset.

##### 4.1.3.1 Create Preprocessing Function

Pada tahapan ini dataset akan melalui beberapa process dengan detail casefolding, tokenizing, stopword, dan stemming.

##### 4.1.3.2 Start Preprocessing

Case folding adalah proses penghilangan karakter selain alfabet. Dengan menggunakan proses ini dataset yang memiliki karakter unik didalamnya akan dihilangkan atau dihapus.

	text	value
0	banyak resto seafoodboga bahari tempat bagus d...	POSITIF
1	pantai kotor	NEGATIF
2	enak buat makan seafood	POSITIF
3	tempat nya bagus hanya saja untuk saat ini mas...	POSITIF
4	aku benci masa orang bayarnya behhhhh bankr...	NEGATIF

Gambar 8. Data Pasca Case Folding

Tokenizing adalah sebuah proses penandaan setiap kata menggunakan token. Setiap kata yang telah ditandai dengan token akan dipisah mennggunakan “,”(koma).

Stopword removal merujuk referensi library untuk menghilangkan kata yang tidak memiliki makna.

Proses stemming adalah proses terakhir dalam melakukan preprocessing penelitian ini. Merubah semua kata kata berimbuhan pada setiap baris dataset menjadi kata dasar. Proses ini mengubah sebuah baris dalam dataset menjadi point penting percakapan.

Dataset yang telah melalui semua proses diatas akan dikumpulkan dan dijadikan sebuah dataset baru.

#### 4.1.4 Feature Extraction and Data Splitting

Feature extraction dilakukan dengan mengubah setiap text menjadi sebuah vector dan angka yang mudah dikenali dengan mesin.

##### 4.1.4.1 Weighting (Vectorize)

Feature extraction ini dilakukan menggunakan pembobotan yang dilakukan oleh TF-IDF.

(0, 2123)	0.30436768740017517
(0, 2196)	0.2220377133298921
(0, 430)	0.3193650200786121
(0, 4369)	0.34738587228546236
(0, 3290)	0.23206673845789558
(0, 926)	0.2251867973034359
(0, 319)	0.17182628325932367
(0, 338)	0.4343346547292049
(0, 4238)	0.4572207016418224
(0, 4019)	0.32187405255317275
(1, 2423)	0.8561271948482095
(1, 3542)	0.5167651557925863
(2, 4236)	0.6200548364229839
(2, 2772)	0.4605598241403215
(2, 1312)	0.6351508861808373
(3, 3528)	0.35405211155738725
(3, 1278)	0.5541885792540818
(3, 2113)	0.42869734764963

Gambar 9. Data Setelah Weighting

#### 4.1.4.2 Data Split

Data splitting dilakukan untuk memisahkan data menjadi dua bagian. Pada kolom x terdapat data training dan data test dan pada kolom y terdapat data training dan data test. Dalam melakukan test data yang digunakan dalam tabel bisa ditentukan melalui `test_size` yaitu menggunakan sample data untuk pengujian keseluruhan dataset.

#### 4.1.5 Modelling

Text classification akan diproses dalam tahap ini. Dengan menggunakan metode Support Vector Machine dataset yang telah melalui semua proses sebelumnya akan dilatih dan dilakukan pengujian dengan kernel 'Linear' pada Probabilitas = True. Program juga akan memprediksi semua data dalam pengujian dengan dibandingkan pada dataset yang telah di *fit*-kan pada data split.

#### 4.1.6 Evaluate

Ditahap ini semua proses pada program akan ditampilkan dengan menarik semua proses pada SVM untuk ditampilkan akurasi, presisi, recall, dan f1 score. Program juga akan memberikan tampilan confusion matrix sebagai gambaran dalam penilaian support karakter.

##### 4.1.6.1 Accuracy, Precision, Recall, F1-Score, Confusion Matrix



Akurasi dalam penelitian ini mencapai 87%, Precision 90%, Recall 65% dan f1score sebesar 75%. Dengan berikut hasil penelitian ini dilakukan menggunakan sample data sebesar 10% dari keseluruhan dataset dan random state 10.

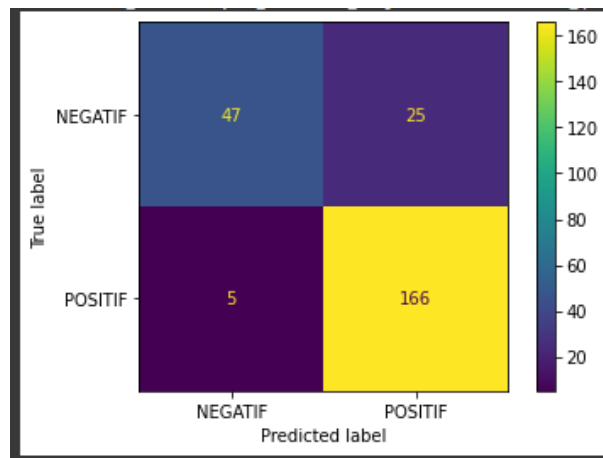
```
SVM Accuracy : 0.8765432098765432
SVM Precision : 0.9038461538461539
SVM Recall : 0.6527777777777778
SVM f1_score : 0.7580645161290323
confusion_matrix :
[[ 47 25]
 [ 5 166]]
```

---

	precision	recall	f1-score	support
NEGATIF	0.90	0.65	0.76	72
POSITIF	0.87	0.97	0.92	171
accuracy			0.88	243
macro avg	0.89	0.81	0.84	243
weighted avg	0.88	0.88	0.87	243

Gambar 10. Evaluasi Model

Pada confusion matrix ditampilkan *True Negative* sebanyak 47, *True Positive* sebanyak 166, *False Positive* sebanyak 25 dan *False Negative* sebanyak 5.



Gambar 11. Confusion Matrix

#### 4.1.6.2 WordCloud

Wordcloud atau bisa disebut sebagai gugus kata akan menampilkan kata kata yang paling sering muncul dalam pengujian dataset baik dengan pengujian data

positive maupun negative. Dengan ini bisa disimpulkan bahwa kata “pantai”, ”Bagus”, “makan” menjadi kata kata yang sering muncul dalam pengujian data positive sedangkan dalam pengujian data negative menampilkan “parkir”, ”pantai”, ”mahal”, ”harga”, “masuk”.



Gambar 12. Wordcloud Positif



Gambar 13. Wordcloud Negatif

#### 4.1.6.3 Cross Validation

Dalam melakukan sebuah penelitian maka akan ada sebuah validasi dalam penelitian tersebut. Dengan menggunakan cross validation yang disetting pada volume sebanyak 10 atau melakukan pengujian model sebanyak 10 kali dataset yang diacak. Penelitian ini menunjukkan bahwa nilai rata – rata cross validation berada pada 83%.

```
Hasil Cross Validation : [0.88477366 0.85596708 0.81069959 0.86419753 0.81404959 0.84297521
0.81404959 0.82644628 0.82231405 0.84297521]
Rata-rata Cross Validation : 0.837844777437676
```

Gambar 14. Hasil Cross Validation

Program ini memberikan sebuah fitur yaitu export dan load model yang berguna apabila program ini akan ditanamkan dalam sebuah software atau website.

#### 4.1.7 Testing

Dalam tahap ini dilakukan percobaan dengan memasukkan kata atau kalimat dengan tujuan menguji model penelitian ini. Dalam bagian testing ini semua proses mulai dari preprocessing hingga pengujian sample akan dilakukan dengan cara mengimport proses tersebut. Dengan menggunakan sample kata “tampatnya bersih dan nyaman” kemudian dimasukkan dalam proses testing akan menampilkan hasil ‘Positive’ dengan Probabilitas 86% yang berarti program memiliki keyakinan sebesar 86% bahwa “bersih dan nyaman” adalah kalimat positive

## **BAB 5. KESIMPULAN**

Projek ini bertujuan untuk mengetahui performa metode SVM dalam klasifikasi analisis sentimen menggunakan metode Support Vector Machine (SVM) berdasarkan ulasan pengunjung pada situs Google Maps. Berdasarkan pengujian yang telah dilakukan, diperoleh nilai rata-rata akurasi sebesar 83.8%.

## DAFTAR PUSTAKA

- [1] “Local Consumer Review Survey: How Customer Reviews Affect Behavior.” <https://www.brightlocal.com/research/local-consumer-review-survey/> (accessed Jan. 10, 2022).
- [2] W. Medhat, A. Hassan, and H. Korashy, “Sentiment analysis algorithms and applications: A survey,” *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, Dec. 2014, doi: 10.1016/J.ASEJ.2014.04.011.
- [3] A. Sasmito Aribowo, H. Basiron, N. Fazilla, A. Yusof, and S. Khomsah, “Cross-domain sentiment analysis model on Indonesian YouTube comment,” *Int. J. Adv. Intell. Informatics*, vol. 7, no. 1, pp. 12–25, 2021, doi: 10.26555/ijain.v7i1.554.
- [4] F. A. Khan, K. Zeb, M. Al-Rakhami, A. Derhab, and S. A. C. Bukhari, “Detection and Prediction of Diabetes Using Data Mining: A Comprehensive Review,” *IEEE Access*, vol. 9, pp. 43711–43735, 2021, doi: 10.1109/ACCESS.2021.3059343.
- [5] P. Wang *et al.*, “Classification of Proactive Personality: Text Mining Based on Weibo Text and Short-Answer Questions Text,” *IEEE Access*, vol. 8, pp. 97370–97382, 2020, doi: 10.1109/ACCESS.2020.2995905.
- [6] N. Zhao, H. Gao, X. Wen, and H. Li, “Combination of convolutional neural network and gated recurrent unit for aspect-based sentiment analysis,” *IEEE Access*, vol. 9, pp. 15561–15569, 2021, doi: 10.1109/ACCESS.2021.3052937.
- [7] C. V D, “Hybrid approach: naive bayes and sentiment VADER for analyzing sentiment of mobile unboxing video comments,” *Int. J. Electr. Comput. Eng.*, vol. 9, no. 5, pp. 4452–4459, 2019, doi: 10.11591/ijece.v9i5.pp4452-4459.
- [8] M. A. Fauzi, “Word2Vec model for sentiment analysis of product reviews in Indonesian language,” *Int. J. Electr. Comput. Eng.*, vol. 9, no. 1, pp. 525–530, 2019, doi: 10.11591/ijece.v9i1.pp525-530.
- [9] A. W. Pradana and M. Hayaty, “The Effect of Stemming and Removal of

Stopwords on the Accuracy of Sentiment Analysis on Indonesian-language Texts,” *Kinet. Game Technol. Inf. Syst. Comput. Network, Comput. Electron. Control*, vol. 4, no. 4, pp. 375–380, 2019, doi: 10.22219/kinetik.v4i4.912.

- [10] E. Sutoyo and A. Almaarif, “Twitter sentiment analysis of the relocation of Indonesia’s capital city,” *Bull. Electr. Eng. Informatics*, vol. 9, no. 4, pp. 1620–1630, 2020, doi: 10.11591/eei.v9i4.2352.
- [11] E. Haddi, X. Liu, and Y. Shi, “The role of text pre-processing in sentiment analysis,” *Procedia Comput. Sci.*, vol. 17, pp. 26–32, 2013, doi: 10.1016/J.PROCS.2013.05.005.
- [12] G. Li and J. Li, “Research on Sentiment Classification for Tang Poetry based on TF-IDF and FP-Growth,” *Proc. 2018 IEEE 3rd Adv. Inf. Technol. Electron. Autom. Control Conf. IAEAC 2018*, pp. 630–634, Dec. 2018, doi: 10.1109/IAEAC.2018.8577715.
- [13] Y. Al Amrani, M. Lazaar, and K. E. El Kadirp, “Random Forest and Support Vector Machine based Hybrid Approach to Sentiment Analysis,” *Procedia Comput. Sci.*, vol. 127, pp. 511–520, Jan. 2018, doi: 10.1016/J.PROCS.2018.01.150.
- [14] T. T. Wong, “Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation,” *Pattern Recognit.*, vol. 48, no. 9, pp. 2839–2846, Sep. 2015, doi: 10.1016/J.PATCOG.2015.03.009.

## **LAMPIRAN**

Lampiran 1 : URL Google Colab

[https://colab.research.google.com/drive/1JyadGIB3ID13OVxjW-Oa\\_9NL-vv7Z3iM?usp=sharing](https://colab.research.google.com/drive/1JyadGIB3ID13OVxjW-Oa_9NL-vv7Z3iM?usp=sharing)

Lampiran 2 : URL Dataset

<https://drive.google.com/drive/folders/1fXJgbLCY1qYI3lh3ANlkYmIkkOnrT6Dq?usp=sharing>