

Analisis Sentimen Terhadap Komentar *Video Youtube* Menggunakan *Support Vector Machines*

Toif Muhayat^{1*}, Ahmad Fauzi², dan Jamaludin Indra³

Program Studi Teknik Informatika, Universitas Buana Perjuangan Karawang, Karawang, Indonesia

*e-mail *Corresponding Author*: if19.toifmuhayat@mhs.ubpkarawang.ac.id

Abstract

Many Youtube users leave comments on the YouTube video content they watch. These comments would be useful if further analysis were carried out. This study aims to analyze Sentiments Toward Youtube Video Comments, to make it easier for creators to find out what types of videos are of interest to viewers, without having to read the comments one by one. The stages carried out in this research are web scrapping, preprocessing, labeling, feature extraction, classification and evaluation. The results of the analysis show that the type of video content with the theme of daily vlogs is more in demand by YouTube users with positive dominant sentiment results. The daily vlog theme has a positive sentiment of 84.0% and a negative sentiment of 16.0%. The use of the SVM (Support Vector Machine) algorithm has an accuracy value of 86%, a precision of 87%, a recall of 99% and an f1-score of 100%.

Keywords: *Sentiment Analysis; Text Mining; Support Vector Machine; Youtube*

Abstrak

Banyak pengguna *Youtube* yang meninggalkan komentar pada konten video *youtube* yang mereka tonton. Komentar-komentar tersebut akan memberikan manfaat jika saja dilakukan analisis lebih lanjut. Penelitian ini bertujuan menganalisis Sentimen Terhadap Komentar Video Youtube, untuk memudahkan kreator mengetahui jenis video yang diminati penonton, tanpa harus membaca komentar secara satu per satu. Tahapan yang dilakukan dalam penelitian ini adalah *web scrapping, preprocessing, labelling, ekstraksi fitur, klasifikasi dan evaluasi*. Hasil analisis menunjukkan jenis konten video bertema *daily vlog* lebih banyak diminati oleh pengguna *youtube* dengan hasil sentimen dominan positif. Tema *daily vlog* memiliki sentimen positif sebesar 84.0% dan sentimen negatif sebesar 16.0%. Penggunaan algoritma SVM (*Support Vector Machine*) memiliki nilai akurasi sebesar 86%, presisi sebesar 87%, *recall* sebesar 99% dan *f1-score* sebesar 100%.

Kata Kunci: *Analisis Sentimen; Text Mining; Support Vector Machine; Youtube*

1. Pendahuluan

Internet pada abad ke 21 dimanfaatkan dalam segala bidang. Mulai dari media sosial, *game online, chatting online, shopping online*, transportasi *online* dan masih banyak lagi. Orang-orang menggunakan internet tidak hanya karena kebutuhan tapi karena ingin mengisi waktu luang. Salah satunya dengan mengunjungi aplikasi *youtube*. Berdasarkan survey yang dilakukan oleh Asosiasi Penyelenggara Jasa Internet Indonesia atau APJII tahun 2022 kuartal 1, 63,02% media sosial yang sering digunakan adalah aplikasi *youtube*.

Setiap pengguna *youtube* dapat membuat *channel* sendiri yang memungkinkan penggunaannya untuk mengunggah video. Pengguna harus membuat konten video yang menarik agar pengguna lain mempunyai rasa ingin tahu dan menonton video yang di-*upload*. Ketika menonton konten video di *youtube*, pengguna merespon video tersebut dengan memberikan bermacam-macam komentar. Beragamnya komentar tekstual yang disampaikan pengguna membuat kreator kesulitan menentukan jenis video yang diminati pengguna. Sehingga perlu adanya analisis sentimen untuk menganalisis komentar yang beragam untuk mencari informasi penting dibalik komentar setiap pelanggan. Analisis sentimen merupakan komputasional dari opini orang lain, pemberian nilai, serta emosi yang terdapat dalam entitas, *event*, dan atribut yang dimiliki [1].

Studi tentang analisis sentimen dapat memberi informasi berharga. Analisis sentimen di jejaring sosial, seperti *twitter*, *youtube* atau *facebook*, telah menjadi sarana yang kuat untuk belajar tentang pendapat pengguna [2]. Analisis sentimen adalah metode menganalisis data untuk mengetahui bagaimana perasaan orang. Analisis sentimen dapat dibagi menjadi 3 tugas, yaitu pengenalan teks informasi, ekstraksi informasi, dan klasifikasi sentimental (deteksi emosi, polaritas) [3].

Text mining digunakan dalam penelitian ini karena telah mendapatkan perhatian yang signifikan di berbagai aplikasi baik di dunia akademis maupun industri. Proyek penelitian dan pertanyaan penelitian menjadi lebih kompleks sehingga yang dibutuhkan lebih dari sekedar pengambilan data. *Text mining* dalam *big data analytics* muncul sebagai alat yang ampuh untuk memanfaatkan kekuatan data tekstual tidak terstruktur dengan menganalisisnya untuk mengekstrak pengetahuan baru untuk mengidentifikasi pola dan korelasi signifikan yang tersembunyi dalam data [4]. Terdapat beberapa metode *text mining* yaitu: *information extraction*, *information retrieval*, *summarization*, *clustering*, *categorization* [5].

Analisis sentimen komentar pengguna youtube dalam penelitian ini bertujuan untuk memudahkan para kreator *youtube* mengetahui jenis video yang diminati pengguna *youtube*, tanpa harus membaca satu persatu komentar yang diberikan. Dataset diambil secara langsung menggunakan *web scraper* di halaman *web youtube*. Selain itu analisis sentimen dapat membantu pengguna *youtube* itu sendiri untuk dapat menikmati konten video yang diinginkan berdasarkan komentar yang dituliskan di *Youtube*. Berdasarkan latar belakang yang telah dipaparkan, algoritme *Support Vector Machine* (SVM) memiliki nilai akurasi yang tinggi dan cukup baik. Maka akan dilakukan penelitian mengenai analisis sentimen terhadap komentar video *youtube* dengan menggunakan algoritma *support vector machines*.

2. Tinjauan Pustaka

Beberapa penelitian sebelumnya mengukur analisis sentimen menggunakan beberapa algoritme misalnya *algoritme SVM*, *Naïve Bayes*, *K-Means* dan *K-Nearest Neighbor*. Penelitian mengenai analisis sentimen pada *twitter* terhadap Nadiem Makariem dengan menggunakan metode *Naïve Bayes* dan *Support Vector Machines* menggunakan data yang diperoleh dari *twitter* melalui *crawling* dengan menggunakan *twitter API*. Hasil uji coba diketahui bahwa sentimen masyarakat terhadap Nadiem Makariem lebih banyak oleh sentimen positif. Hasil penelitian menunjukkan bahwa algoritme *Naïve Bayes* memiliki akurasi lebih besar dari SVM. *Naïve Bayes* mempunyai akurasi 91.48% sedangkan SVM memiliki keakurasian 85.47% [6].

Penelitian lainnya yaitu tentang sentimen analisis dan klasifikasi *tweets* yang menggunakan bahasa Indonesia terhadap moda transportasi umum yaitu Moda Raya Terpadu (MRT) Jakarta dengan menggunakan algoritme *Naïve Bayes*. Data diambil dari *twitter* melalui proses *crawling*, hasil penelitiannya adalah akurasi sistem yang diperoleh dari analisis sentimen terhadap *tweet* mengenai MRT Jakarta adalah 95.88% [7].

Penelitian tentang analisis sentimen dan klasifikasi terhadap berita *hoax* Covid-19 dengan menggunakan algoritme *Support Vector Machines*. Berdasarkan hasil dari penelitian ini, algoritme SVM dengan *kernel linear* menunjukkan hasil prediksi yang baik pada skenario (80:20). Pasalnya, model tersebut mampu membedakan berita yang *hoax* dan bukan *hoax* tentang Covid-19 dengan akurasi tertinggi yaitu 97,06%. Kemudian hasil akurasi *kernel RBF* paling rendah pada skenario (90:10) yaitu 90,46%, model ini tidak cocok untuk mengklasifikasikan berita *hoax* atau tidak *hoax* Covid-19 [8].

Pada tahun 2022 penelitian dilakukan oleh Miranti, Nurul, Muhammad dan Galuh mengenai analisis sentimen *twitter* menggunakan data *tweet twitter* dengan 1000 data, yang bertujuan untuk mengetahui keberadaan *K-poppers* di Indonesia. Hasil penelitian menunjukkan bahwa sentimen negatif sebesar 15,09%, netral sebesar 51,75%, dan positif sebesar 33,15%. Tingkat evaluasi yang diperoleh dengan menggunakan metode *Silhouette Coefficient* adalah sebesar 0.687974 [9].

Penelitian yang dilakukan oleh Auliya, Adam, Debby, dan Nurman mengenai analisis sentimen terhadap kebijakan *lock down* pemerintah Jakarta menggunakan algoritme SVM dengan ekstraksi fitur *tf-idf*. Data diambil melalui proses *crawling* pada *twitter*, hasil klasifikasi mendapatkan sentimen positif sebesar 68,75% dan negatif sebesar 31,25%. Nilai akurasi dalam menggunakan algoritme SVM adalah sebesar 74%, nilai presisi sebesar 75%, nilai *recall* sebesar 92% dan nilai sebesar *f1-score* 83% [10].

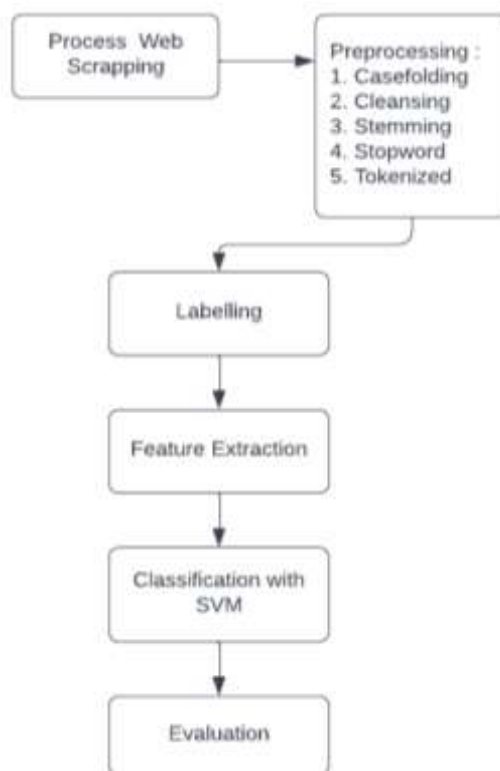
Penelitian berjudul "Analisis Sentimen Aplikasi Ruang Guru Di Twitter Menggunakan Algoritme Klasifikasi" melakukan analisis sentimen terhadap aplikasi Ruang Guru dengan data set diambil dari *twitter*. Dilakukan klasifikasi menggunakan algoritme *Naive Bayes* (NB), *Support*

Vector Machine (SVM), *K-Nearest Neighbour* (K-NN). Hasil pengujian menunjukkan bahwa model terbaik adalah algoritme PSO berbasis SVM dengan nilai akurasi 78,55% dan AUC sebesar 0,853 [11].

Penelitian yang membahas analisis sentimen sudah banyak dilakukan. Analisis sentimen yang dilakukan sudah menggunakan algoritme yang berbeda-beda, misalnya SVM, *Naïve Bayes*, *K-Means* dan *K-Nearest Neighbor* dan dataset yang diambil adalah menggunakan data twitter. Pada penelitian ini menggunakan algoritme SVM dengan mengambil dataset dari *youtube* dengan tujuan dapat mengetahui tema konten video *youtube* yang diminati oleh pengguna *youtube*. Pengambilan dataset dari aplikasi *youtube* inilah yang membedakan penelitian ini dengan penelitian yang lain yang menggunakan dataset dari twitter. Dataset diambil dari komentar yang diberikan oleh pengguna pada konten video di *youtube*. Penggunaan algoritme *support vector machines kernel polynomial* dengan bahasa pemrograman *python* diharapkan dapat mengklasifikasi komentar video *youtube* dengan baik, sehingga dapat diketahui jenis konten video dengan sentimen dominan positif dan konten video dengan komentar dominan negatif.

3. Metodologi

Terdapat beberapa tahapan yang dilaksanakan dalam penelitian ini dapat dilihat pada Gambar 1.



Gambar 1. Diagram Alir Tahapan Penelitian

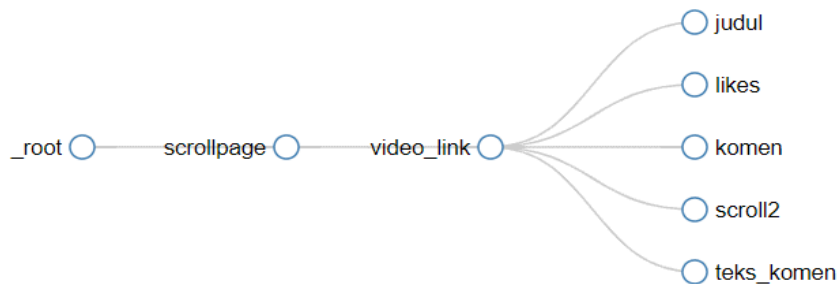
3.1. Web Scrapping

Teknik Web Scraping digunakan untuk teknik pengambilan data suatu web. Metode *Depth First Search* (DFS) merupakan metode pencarian yang bisa digunakan dalam proses *Web Scraping* di Internet [12]. *Web scraping* adalah sebuah tahapan pengambilan data yang bertindak sebagai pengumpul dataset yang merupakan langkah pertama dari penelitian ini. Data yang diambil berupa komentar pada *channel youtube* Ricis Official dengan tema *daily vlog*, *channel*

Ken & Grat dengan tema video kuliner, *channel* GadgetIn dengan tema video *unboxing*, *channel* 123 GO! Indonesian dengan tema video hack / DIY, dan *channel* Tereza Fahlevi dengan tema video *cover music*. *Selector graph* digunakan sebagai alur dari *web scrapper* dalam mengambil data. Gambar 2 merupakan *selector graph* yang digunakan dalam proses *Web Scrapping* pada penelitian ini.

3.2. Preprocessing

Preprocessing data melibatkan pembersihan data, sehingga data dapat digunakan untuk tahap selanjutnya. Tahapan *preprocessing* yang dilakukan dalam penelitian ini meliputi *case*



Gambar 2. Selector Graph

folding, *cleansing*, *stemming*, *stopwords*, dan *tokenization*.

- Case folding*, proses mengubah teks dokumen menjadi huruf
- Cleansing*, proses membersihkan teks dokumen dari teks yang tidak dibutuhkan oleh proses seperti, *hashtag*, *html*, *emoticon*, *url* dan *mention*
- Stopwords*, proses menghapus kata sambung yang tidak efektif
- Stemming*, bertujuan untuk mengubah kata pada teks dokumen menjadi kata dasar
- Tokenization*, merupakan proses memotong setiap kata dalam kalimat dengan pemisah tanda petik (').

3.3. Labelling

Proses *labelling* dilakukan menggunakan kamus bahasa *lexicon* dengan mengosongkan kamus bahasa *lexicon default* terlebih dahulu, kemudian mengisi kamus *lexicon* menggunakan kamus sentimen berbahasa indonesia oleh Wahid [13].

3.4. Ekstraksi Fitur

Sebelum dilakukan klasifikasi terlebih dahulu dilakukan pembobotan setiap kata menggunakan *TfidfVectorizer* dari *library sklearn* [14]. Dengan menggunakan ekstraksi fitur TF-IDF akan meningkatkan efisiensi klasifikasi yang akan mengurangi waktu pengolahan oleh komputer dan meningkatkan akurasi.

3.5. Klasifikasi

Dalam melakukan pelatihan dengan algoritme SVM mengikuti aturan sebagai berikut:

1. Titik data:

$$x_i = \{x_1, x_2, \dots, x_n\} \in R^n \dots\dots(5)$$

2. Kelas data:

$$y_i \in \{-1, +1\} \dots\dots(6)$$

3. Pasangan data dan kelas :

$$\{(x_i, y_i)\}_{i=1}^N \dots\dots(7)$$

4. Maksimalkan fungsi berikut:

$$Ld = \sum_{i=1}^N \alpha_i - \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j K(x_i, x_j) \dots\dots(8)$$

$$\text{syarat} : 0 \leq \alpha_i \leq C \text{ dan } \sum_{i=1}^N \alpha_i y_i = 0 \dots\dots(9)$$

5. Hitung nilai bobot *vector* (w) dan bias (b):

$$w = \sum_{i=1}^N \alpha_i y_i x_i \dots\dots(10)$$

$$b = -\frac{1}{2}(w \cdot x^+ + w \cdot x^-) \dots\dots(11)$$

6. Fungsi keputusan klasifikasi *sign* (f(x)) :

$$f(x) = w \cdot x + b \dots\dots(12)$$

atau

$$f(x) = \sum_{i=1}^m \alpha_i y_i K(x, x_i) + b \dots\dots(13)$$

Keterangan:

N : Jumlah data yang ada

n : Jumlah fitur yang ada

Ld : Dualitas *Lagrange Multiplier*

α_i : Nilai bobot setiap titik data

C : Nilai konstanta

M : Jumlah *support vector*/titik data yang memiliki $\alpha_i > 0$

$K(x, x_i)$: Fungsi *Kernel* [15].

Sebuah *support vector machine* mengambil dua titik (vektor) dan membentuk garis pemisah (atau permukaan batas jika ada lebih dari dua dimensi). Garis pemisah yang dibentuk oleh kedua vektor ini disebut *hyperplane* [10]. *Kernel* yang digunakan dalam penelitian ini adalah *kernel polynomial*. Tabel 2 merupakan rumus *kernel* pada *support vector machine*.

Tabel 2. Rumus Kernel

No	Jenis Kernel	Model
1	<i>Linear</i>	$K(x, x') = (x \cdot x') \dots\dots(1)$
2	<i>Polynomial</i>	$K(x, x') = (x \cdot x' + c)^n \dots\dots(2)$
3	<i>RBF Gaussian</i>	$K(x, x') = \exp(-\gamma x - x' ^2) \dots\dots(3)$
4	<i>Sigmoid</i>	$K(x, x') = \tanh(\alpha x \cdot x' + \beta) \dots\dots(4)$

3.6. Evaluasi

Tahap evaluasi berfungsi untuk menghitung nilai akurasi, *recall* dan presisi. Metode *Confusion Matrix* digunakan untuk melakukan evaluasi dengan membandingkan hasil pelabelan dengan hasil klasifikasi SVM. Tabel 3 di bawah merupakan perhitungan *Confusion Matrix*.

Tabel 3. Confusion Matrix

		Hasil Klasifikasi	
		Positif	Negatif
Kelas Actual	Positif	TP	FN
	Negatif	FP	TN

Akurasi merupakan hasil prediksi positif dan negatif yang dibandingkan dengan keseluruhan data yang diuji. Persamaan (14) merupakan rumus untuk menghitung akurasi.

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \dots\dots(14)$$

Recall merupakan bagian kelas positif yang diklasifikasi secara tepat. Persamaan (15) adalah rumus untuk menghitung nilai *recall*.

$$Recall = \frac{TP}{TP + FN} \dots\dots(15)$$

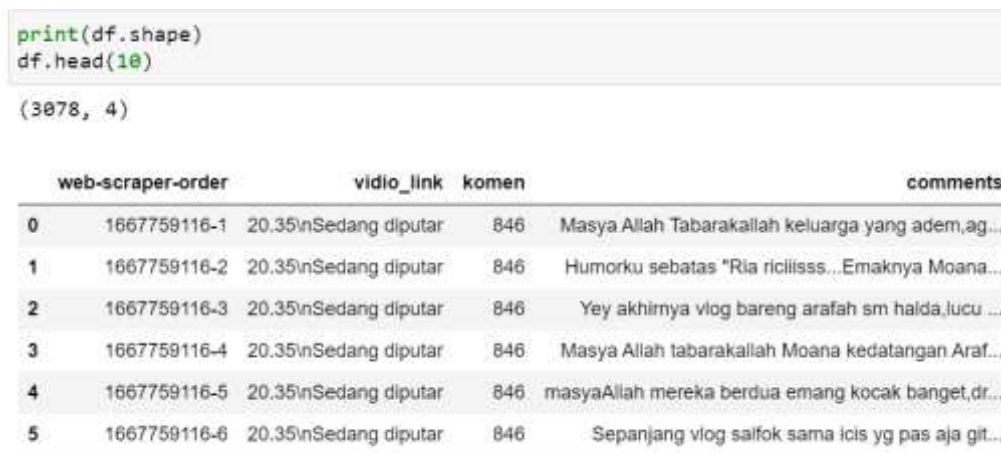
Presisi adalah proporsi kelas positif yang diklasifikasikan sebagai benar positif dibandingkan dengan jumlah total hasil yang diklasifikasikan sebagai positif. Berikut persamaan (16) untuk menghitung presisi.

$$\text{Presisi} = \frac{TP}{TP + FP} \dots\dots(16)$$

4. Hasil dan Pembahasan

4.1. Web Scrapping

Proses *web scrapping* dilakukan tanggal 07 November 2022 pada akun *youtube* Ricis Official, Ken & Grat, GadgetIn, 123 GO! Indonesia, Tereza Fahlevi. Dengan contoh hasil *web scrapping channel youtube* Ricis Official yang terdapat pada Gambar 3.



```
print(df.shape)
df.head(10)

(3978, 4)
```

	web-scraper-order	vidio_link	komen	comments
0	1667759116-1	20.35\nSedang diputar	846	Masya Allah Tabarakallah keluarga yang adem,ag...
1	1667759116-2	20.35\nSedang diputar	846	Humorku sebatas "Ria riciiss...Emaknya Moana...
2	1667759116-3	20.35\nSedang diputar	846	Yey akhirnya vlog bareng arafah sm halda,lucu ...
3	1667759116-4	20.35\nSedang diputar	846	Masya Allah tabarakallah Moana kedatangan Araf...
4	1667759116-5	20.35\nSedang diputar	846	masyaAllah mereka berdua emang kocak banget,dr...
5	1667759116-6	20.35\nSedang diputar	846	Sepanjang vlog saifok sama icis yg pas aja git...

Gambar 3. Hasil *Web Scrapping*

4.2. Preprocessing

Case folding, cleansing text, stemming, stopword, tokenization, dilakukan pada fase ini. Proses pada tahap ini yaitu mengonversi semua karakter dalam dokumen teks menjadi huruf kecil. Selanjutnya menghapus tanda baca dan karakter asing titik, tanda tanya, tanda seru, koma, *url, html*, emoji, dan menghapus simbol asing, menghapus kata henti, mengubah kata majemuk menjadi kata dasar dan kemudian memisahkan teks.

4.2.1. Case Folding

Pada tahap ini untuk mengubah data teks menjadi huruf kecil (*lower case*) agar mudah dalam pembacaan komputer. Tabel 4 merupakan kutipan hasil yang diperoleh dari tahap *case folding*.

Tabel 4. Hasil *Case Folding*

Sebelum	Sesudah
Masya Allah Tabarakallah keluarga yg adem,agamis dan harmonis yang menjadi panutan semua orang,semoga selalu sehat ² .	masya allah tabarakallah keluarga yg adem,agamis dan harmonis yang menjadi panutan semua orang,semoga selalu sehat ² .
Humorku sebatas "Ria riciiss...Emaknya Moana, Ria Ricis istrinya pak Tengku" Astagfirullah	humorku sebatas "ria riciiss...emaknya moana, ria ricis istrinya pak tengku" astagfirullah

4.2.2. Cleansing Text

Pada tahap ini yaitu menghapus tautan *url*, titik, koma, tanda seru, tanda tanya, penghapusan emoji, sebutan dan tagar, dan hapus simbol yang tidak penting. Tabel 5 merupakan kutipan hasil dari proses *cleansing text*.

Tabel 5. Hasil *Cleansing Text*

Sebelum	Sesudah
masya allah tabarakallah keluarga yg adem,agamis dan harmonis yang menjadi panutan semua orang,semoga selalu sehat ² .	masya allah tabarakallah keluarga yg adem agamis dan harmonis yang menjadi panutan semua orang semoga selalu sehat
humorku sebatas "ria riciiss...emaknya moana, ria ricis istrinya pak tengku" astagfirullah	humorku sebatas ria ricis emaknya moana ria ricis istrinya pak tengku astagfirullah

4.2.3. Stemming

Tujuan dilakukan *stemming* untuk menemukan kata dasar, *library stemmer factory* digunakan dalam penelitian ini. Tabel 6 merupakan kutipan hasil dari proses *stemming*.

Tabel 6. Hasil *Stemming*

Sebelum	Sesudah
masya allah tabarakallah keluarga yg adem agamis dan harmonis yang menjadi panutan semua orang semoga selalu sehat	masya allah tabarakallah keluarga yg adem agama harmonis panutan orang moga sehat
humorku sebatas ria ricis emaknya moana ria ricis istrinya pak tengku astagfirullah	humor batas ria ricis emaknya moana ria ricis istri tengku astagfirullah

4.2.4. Stopword

Untuk mempermudah proses *remove stopwords*, digunakan *library NLTK* berbahasa indonesia. Penelitian ini menambahkan beberapa kata yang sering dijumpai pada data seperti "tdk", "yg", "utk" dan lainnya untuk meminimalkan adanya *noise* dan membuat data lebih bersih. Tabel 7 merupakan kutipan hasil dari proses *stopword*.

Tabel 7. Hasil *Stopword*

Sebelum	Sesudah
masya allah tabarakallah keluarga yg adem agama harmonis panutan orang moga sehat	masya allah tabarakallah keluarga adem agama harmonis panutan orang moga sehat
humor batas ria ricis emaknya moana ria ricis istri tengku astagfirullah	humor batas ria ricis emaknya moana ria ricis istri tengku astagfirullah

4.2.5. Tokenization

Tokenization bertujuan untuk memecah teks menjadi beberapa bagian, dapat berupa kata atau kalimat. Proses ini menggunakan *library NLTK* khusus untuk *tokenization*. Tabel 8 merupakan kutipan hasil dari proses *tokenization*.

Tabel 8. Hasil *Tokenization*

Sebelum	Sesudah
masya allah tabarakallah keluarga adem agama harmonis panutan orang moga sehat	['masya', 'allah', 'tabarakallah', 'keluarga', 'adem', 'agama', 'harmonis', 'panutan', 'orang', 'moga', 'sehat', '']
humor batas ria ricis emaknya moana ria ricis istri tengku astagfirullah	['humor', 'batas', 'ria', 'ricis', 'emaknya', 'moana', 'ria', 'ricis', 'istri', 'tengku', 'astagfirullah']

4.3 Labelling

Data sentimen komentar video *youtube* dilakukan proses *labelling* menggunakan kamus sentimen *lexicon* yang dikosongkan. Proses selanjutnya adalah pengisian *lexicon* dengan kamus sentimen bahasa indonesia dari Wahid untuk membuat label sentimen negatif dan positif. Tabel 9 merupakan cuplikan hasil dari proses *labelling*.

Tabel 9. Hasil *Labelling*

Komentar	Sentimen
masyaallah tabarakallah sehat dan sukses buat ricis ryan dan baby moana dan seluruh teamnya smoga dijauhkan dari fitnah dan orang yang iri dengki dengan kebahagiaan keharmonisan kluarga ricis aamiin	Negatif
liat kalian berdua seneng banget dehsehat selali kalian berdua smoga rumah tangga kalian di jauhkan dari perselingkuhan perselisihan aminn	Negatif
masya allah imut banget loh moana lucu banget	Positif
masya allah semoga moana jadi anak yang pintar baik dan solehah amin	Positif

4.4 Ekstraksi Fitur

Bertujuan untuk meningkatkan efisiensi *klasifikasi*, meningkatkan akurasi dan mengurangi waktu komputasi. Gambar 4 merupakan hasil proses TF-IDF.

```
[ 'aa', 'aah', 'aahh', 'aalemin', 'aamii', 'aamin', 'aaminn', 'aamiinsehat',
  [ 'yunifa', 'yunifah', 'yunita', 'yusuf', 'yutub', 'yutuber', 'yy', 'zaman', '
[[0. 0. 0. 0. 0. 0. ...
 0. 0. 0. 0. 0. 0. ]
 0. 0. 0. 0. 0. 0. ...
 0. 0. 0. 0. 0. 0. ]
 0. 0. 0. 0. 0. 0.06379658 ...
 0. 0. 0. 0. 0. 0. ]
 0. 0. 0. 0. 0. 0.11743204 ...
 0. 0. 0. 0. 0. 0. ]
 0. 0. 0. 0. 0. 0. ...
 0. 0. 0. 0. 0. 0.08438629 ...
 0. 0. 0. 0. 0. 0. ]
 0. 0. 0. 0. 0. 0. ...
 0. 0. 0. 0. 0. 0. ]]
{'aa': 0, 'aah': 1, 'aahh': 2, 'aalemin': 3, ...}
['masya', 'allah', 'tabarakallah', 'keluarga', 'yang', 'aden', 'agamis', 'dan'
```

Gambar 4. Hasil TF-IDF

4.5 Klasifikasi

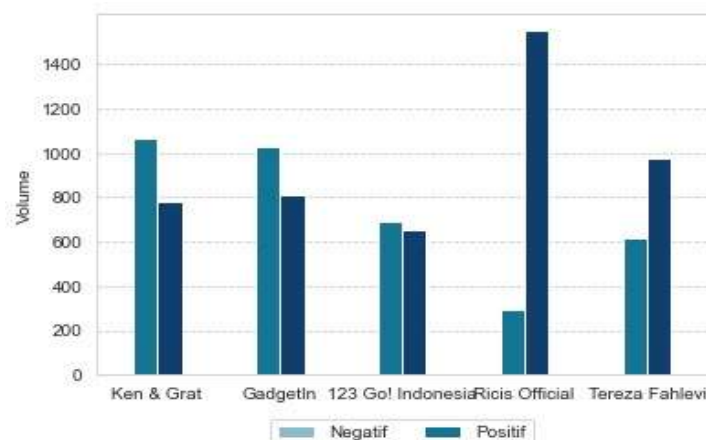
Pada klasifikasi dilakukan *split data training* dan *testing*, yaitu 90% sebagai data *training* kemudian 10% sebagai *random testing*. Tabel 10 merupakan hasil klasifikasi menggunakan *support vector machines*.

Tabel 10. Hasil Klasifikasi Support Vector Machines

Channel Youtube	Sentimen Positif	Sentimen Negatif
Ricis Official	84.0%	16.0%
Ken & Grat	42.2%	57.8%
GadgetIn	42.2%	57.8%
123GO! Indonesian	48.7%	51.3%
Tereza Fahlevi	61.4%	38.6%

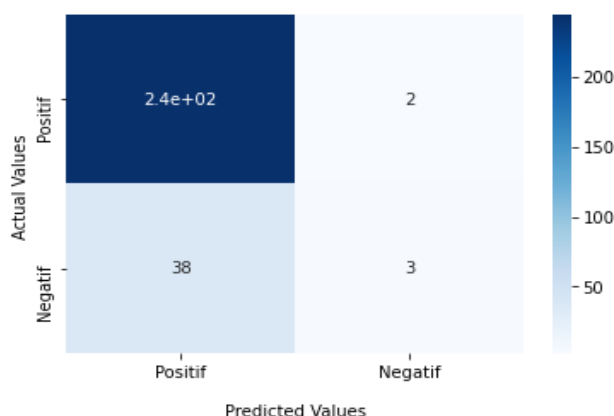
Berdasarkan hasil klasifikasi, *channel youtube* Ricis Official memiliki sentimen positif tertinggi yaitu 84.0%, dan sentimen negatif sebesar 16.0%, sedangkan *channel youtube* Ken & Grat dan Gadgetin memiliki sentimen positif terendah yaitu 42.2% dan sentimen negatif sebesar 57.8%. Gambar 5 merupakan Visualisasi data dari hasil klasifikasi support vector machines.

Channel youtube Ricis Official memiliki kemunculan kata dengan frekuensi paling besar pada komentar yaitu kata Moana. Gambar 6 merupakan contoh kemunculan kata terbesar pada *channel youtube* Ricis Official.



4.6 Evaluasi

Pengujian menggunakan metode *confusion matrix* dilakukan untuk mendapatkan nilai akurasi, presisi, *recall* dan *f1-score*. Gambar 7 merupakan hasil yang diperoleh dari pengujian *confusion matrix*.



Gambar 7. Confusion Matrix

Confusion matrix pada Gambar 7 diatas adalah hasil pengujian *support vector machine*, dengan pembagian data 90:10 dimana terdapat nilai *true positive* yaitu ketika data aktual positif dan berhasil diprediksi sebagai positif, kemudiann nilai *true negative* yaitu ketika data aktual negatif dan berhasil diprediksi sebagai negatif, kemudian *false positive* yaitu ketika data aktual negatif dan diprediksi sebagai positif, sedangkan *false negative* yaitu ketika data aktual positif dan diprediksi sebagai negatif.

Hasil *confusion matrix* pada Gambar 7 diatas nilai *true positive* sebesar 245 data, nilai *true negative* 3 data, nilai *false positive* 38 data dan *false negative* 2 data, menghasilkan akurasi *support vector machine* sebesar 86%, presisi sebesar 87%, *recall* sebesar 99% dan *f1-score* sebesar 100%. Berdasarkan akurasi yang diperoleh yaitu sebesar 86% dengan menggunakan *support vector machine kernel polynomial*, maka penelitian mengenai analisis sentimen komentar video youtube dikatakan cukup baik.

Berdasarkan hasil penelitian yang telah dilakukan bahwa klasifikasi sentimen komentar pengguna youtube menggunakan *support vector machine kernel polynomial* menghasilkan akurasi sebesar 86%. Hal ini berarti SVM dapat membantu kreator untuk mengetahui jenis video yang diminati oleh pengguna youtube dengan cukup baik. Hasil akurasi klasifikasi sentimen menggunakan *support vector machine* yang baik ini menguatkan penelitian terdahulu dengan hasil yang diperoleh pada penelitian terdahulu yaitu penelitian [6] nilai akurasi sebesar 85.47% dengan data berjumlah 2500 data dan penelitian [8] nilai akurasi sebesar 90.46% dengan data berjumlah 1377 data, sedangkan penelitian ini menggunakan data yang berjumlah 3078 data.

Perbedaan penelitian ini dengan penelitian terdahulu yaitu pada pada penelitian ini data diambil dari komentar video youtube dan penelitian terdahulu menggunakan data teks dari *tweet* pada *twitter*. Walaupun jumlah dan sumber data yang digunakan berbeda namun hasil akhir dari penelitian ini memperkuat penelitian [6], [8] bahwa SVM cukup baik dalam melakukan klasifikasi sentimen. Penelitian selanjutnya dapat menggunakan metode seleksi fitur lainnya untuk mendapatkan hasil akurasi yang lebih maksimal. Dapat juga dibuatkan aplikasi berbasis web untuk memudahkan analisis sentimen dengan dataset yang lebih beragam.

5. Simpulan

Klasifikasi analisis sentimen dilakukan terhadap komentar video youtube berbahasa Indonesia dengan tema *daily vlog*, video kuliner, video *unboxing*, video *hack/DIY*, video *cover music*. Kesimpulan jenis konten video youtube dengan sentimen dominan positif yaitu jenis konten *daily vlog*. Presentase sentimen positif yang diperoleh konten *daily vlog* sebesar 84.0% dan presentase sentimen negatif sebesar 16.0%. Sedangkan jenis konten video youtube dengan sentimen dominan negatif adalah jenis konten *unboxing* dan jenis konten video kuliner dengan sentimen positif 42.2% dan sentimen negatif 57.8%. Berdasarkan hasil tersebut, sentimen ini bisa digunakan sebagai pertimbangan dalam pembuatan jenis konten video youtube. Algoritme *support vector machines* menggunakan ekstraksi fitur *tf-idf* pada klasifikasi komentar video youtube berbahasa Indonesia, menggunakan *kernel polynomial* memiliki nilai akurasi sebesar 86%, presisi sebesar 87%, *recall* sebesar 99% dan *f1-score* sebesar 100%. Berdasarkan hasil penelitian yang telah dilakukan maka dapat dikatakan bahwa penggunaan algoritme *support vector machines* cukup baik pada penelitian ini.

Daftar Referensi

- [1] F. A. Nugraha, N. H. Harani, and R. Habibi, *Analisis Sentimen Pembatasan Sosial Menggunakan Deep Learning*, 1st ed. Bandung: Kreatif, 2020.
- [2] N. C. Dang, M. N. Moreno-García, and F. de la Prieta, "Sentiment analysis based on deep learning: A comparative study," *Electronics (Switzerland)*, vol. 9, no. 3, Mar. 2020, doi: 10.3390/electronics9030483.
- [3] A. M. Siregar and T. A. Hasan, "Aplikasi Linier Regresi Dengan Algoritma Jaringan Syaraf Tiruan Untuk Sentimen Analisis," *Techno Xplore: Jurnal Ilmu Komputer dan Teknologi Informasi*, vol. 3, no. 2, pp. 43-51, Oct. 2018.
- [4] H. Hassani, C. Beneki, S. Unger, M. T. Mazinani, and M. R. Yeganegi, "Text mining in big data analytics," *Big Data and Cognitive Computing*, vol. 4, no. 1, pp. 1–34, Mar. 2020, doi: 10.3390/bdcc4010001.
- [5] S. Tandel, "A Survey on Text Mining Techniques," *International Conference on Advanced Computing & Communication Systems*, pp. 1–5, 2019.
- [6] T. T. Widowati and M. Sadikin, "Analisis Sentimen Twitter Terhadap Tokoh Publik dengan Algoritme Naïve Bayes dan Support Vector Machine," *Jurnal SIMETRIS*, vol. 11, no. 2, 2020, [Online]. Available: <https://t.co/Xzf91zHK41>
- [7] D. I. Sari, Y. F. Wati, and Widiastuti, "Analisis Sentimen dan Klasifikasi Tweets Berbahasa Indonesia Terhadap Transportasi Umum MRT Jakarta Menggunakan Naive Bayes Classifier," *Jurnal Ilmiah Informatika Komputer*, vol. 25, no. 1, pp. 64–75, 2020, doi: 10.35760/ik.2020.v25i1.2427.
- [8] I. A. Ropikoh, R. Abdulhakim, U. Enri, and N. Sulistiyowati, "Penerapan Algoritme Support Vector Machine (SVM) untuk Klasifikasi Berita Hoax Covid-19," 2021. [Online]. Available: <http://jurnal.polibatam.ac.id/index.php/JAIC>
- [9] M. A. Z. Larasati, N. A. S. Winarsih, M. S. Rohman, and G. W. Saraswati, "Penerapan Metode K-Means Clustering dalam Menganalisis Sentimen Masyarakat terhadap K-Popers pada Twitter," *Progresif: Jurnal Ilmiah Komputer*, vol. 18, no. 2, pp. 201–210, Aug. 2022, [Online]. Available: <https://databoks.katadata.co.id/>
- [10] A. Rahman Isnain, A. Indra Sakti, D. Alita, and N. Satya Marga, "Sentimen Analisis Publik Terhadap Kebijakan Lockdown Pemerintah Jakarta Menggunakan Algoritme SVM," *JDMSI*, vol. 2, no. 1, pp. 31–37, 2021, [Online]. Available: <https://t.co/NfhnfMjtXw>
- [11] A. P. Giovani, A. Ardiansyah, T. Haryanti, L. Kurniawati, and W. Gata, "Analisis Sentimen Aplikasi Ruang Guru Di Twitter Menggunakan Algoritme Klasifikasi," *Jurnal Teknoinfo*, vol. 14, no. 2, p. 115, Jul. 2020, doi: 10.33365/jti.v14i2.679.
- [12] V. A. Flores, P. A. Permatasari, and L. Jasa, "Penerapan Web Scraping Sebagai Media Pencarian dan Menyimpan Artikel Ilmiah Secara Otomatis Berdasarkan Keyword," *Majalah Ilmiah Teknologi Elektro*, vol. 19, no. 2, p. 157, Dec. 2020, doi: 10.24843/mite.2020.v19i02.p06.
- [13] D. H. Wahid and S. N. Azhari, "Peringkasan Sentimen Esktraktif di Twitter Menggunakan Hybrid TF-IDF dan Cosine Similarity," *Indonesian Journal of Computing and Cybernetics Systems*, pp. 207–218, 2016.
- [14] H. Sujadi, "Analisis Sentimen Pengguna Media Sosial Twitter Terhadap Wabah Covid-19 Dengan Metode Naive Bayes Classifier Dan Support Vector Machine," *INFOTECH journal*, vol. 8, no. 1, pp. 22–27, Mar. 2022, doi: 10.31949/infotech.v8i1.1883.
- [15] F. A. Sianturi, P. M. Hasugian, A. Simangunsong, and B. Nadeak, *Data Mining Teori dan Aplikasi Weka*. IOCS Publisher, 2019.