

# Analisis Sentimen LPDP (Lembaga Pengelola Dana Pendidikan) Menggunakan SVM dan *Naïve Bayes* Pada Media Sosial Twitter

Samuel Adhi Bagaskoro<sup>a</sup>, Atin Hasanah<sup>b\*</sup>, Saiful Bahri<sup>c</sup>, Ema Utami<sup>d</sup>, Ainul Yaqin<sup>e</sup>

<sup>a,b,c,d,e</sup>Universitas AMIKOM Yogyakarta, Yogyakarta 55283, Indonesia

Informasi Naskah:

Diterima: 31 Juli 2022 /Direview: 04 Juli 2023/ Direvisi: 13 Juli 2023 / Disetujui Terbit: 19 Juli 2023

DOI: <https://doi.org/10.33369/pseudocode.10.2.65-73>

\*Korespondensi: atin.hasanah92s2@students.amikom.ac.id

## Abstract

*LPDP Scholarship (Education Fund Management Institution) is the most sought after by prospective students who want to continue their studies in the country, especially for those who want to continue their studies abroad. Recently, LPDP experienced problems related to students who received LPDP scholarships but did not return to Indonesia in accordance with the rules that have been stated. Starting from the incident on twitter, the topic of "LPDP" became a trending topic among twitter users. So it is our concern to find out and analyze public opinion through this twitter social media. By comparing the results of two methods, namely Support Vector Machine (SVM) and Naïve Bayes in classifying the twitter sentiment. As well as the calculation of accuracy using the Confusion Matrix, there are as many as 1000 tweets result from crawling. This research resulted in a classification that uses the Vader Lexicon Library built by NLTK, the Naïve Bayes method and Support Vector Machine (SVM) has not yet reached an accuracy rate of 70%. In contrast, the Support Vector Machine (SVM) method that uses the Vader Lexicon Library from VaderSentiment achieves an accuracy rate of 90%, with a ratio of 90:10 (training data: test data).*

*Keywords: LPDP, Naïve Bayes, Sentiment Analysis, Support Vector Machine (SVM), Vader Lexicon.*

## 1. Pendahuluan

Beasiswa LPDP (Lembaga Pengelola Dana Pendidikan) yang paling diburu para calon mahasiswa yang ingin melanjutkan kuliahnya di dalam negeri terutama bagi yang ingin meneruskan kuliah ke luar negeri [1]. Berbagai persyaratan harus terpenuhi untuk dapat lulus administrasi pendaftaran LPDP ini. Selain itu, para calon mahasiswa tersebut harus mengikuti tes sebagai tahap lanjutan dan terakhir merupakan kegiatan diklat yang diselenggarakan langsung oleh pihak LPDP.

Baru-baru ini, LPDP mengalami permasalahan berkaitan dengan para mahasiswa yang menerima beasiswa LPDP namun tidak kembali ke Indonesia sesuai dengan aturan yang sudah tertera. Seperti yang dilansir oleh web <https://www.suara.com/> bahwa ada akun twitter yang mengungkap kasus tentang adanya penerima beasiswa luar negeri yang tidak pulang ke Indonesia.

Berawal dari kejadian tersebut di twitter, topik "LPDP" menjadi trending topik di kalangan pengguna twitter. Disana semakin banyak pengguna twitter yang mengungkap kejadian-kejadian tersebut. Tercatat sebanyak 11.300 tweet per tanggal 29 Juli 2022, sehingga ini menjadi perhatian kami untuk mencari tahu dan menganalisis opini masyarakat melalui media sosial twitter ini.

Dalam penelitian Santoso dkk, yang diusulkan menggunakan algoritme *detection* dengan metode *searching* dengan beberapa tahap, user menginputkan berita *hoax* selanjutnya dilakukan klasifikasi menggunakan algoritme

pencarian, dari hasil klasifikasi tersebut menggunakan algoritme cosinus similarity dengan pengujian 30 berita, ditemukan rata-rata 77% berita *hoax* dengan presentase minimal memiliki nilai 66% untuk presentase maksimal 91% [2].

Arsi dan Waluyo mengusulkan metode SVM dengan tujuan mengklasifikasi kelas sentimen pada twitter menjadi kelas positif 404 dan negatif 832, dari hasil uji pada twitter sebanyak 1.236 tweets dengan akurasi mencapai 96,68%, presisi 95,82%, recall 94,04% dan AUC 0,979 [3].

Lalu Aljameel dkk, tersebut mengklasifikasi menggunakan metode SVM dan biagram dalam Term Frequency-Inverse dan TF-IDF memiliki akurasi yang cukup tinggi yaitu 85% dalam kesadaran tertinggi untuk melawan COVID-19 [4].

Metode *Naïve Bayes* pada penelitian Rifa'I dkk untuk mengklasifikasi kelas sentimen pada tempat wisata menggunakan Google Maps. Hal tersebut memiliki akurasi yang bervariasi pada objek wisata dengan nilai akurasi maksimal mencapai 0,76% dan minimal 0,38%. Hasil analisis pada tempat wisata yang berlokasi di kalimantan barat memiliki kategori rata-rata f1-score 0,73% (positif), 0,53% (netral) dan 0,14% (negatif) [5].

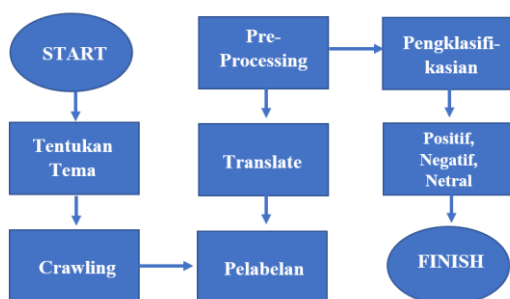
Maka dari itu, peneliti ingin melakukan analisis *sentiment* terkait LPDP yang selalu diburu oleh para mahasiswa namun mahasiswa yang memperolehnya tidak kembali ke Indonesia ketika mereka sudah lulus kuliah. Selain itu, terungkap kasus-kasus terkait hal tersebut sehingga menjadi topik hangat di twitter. Sehingga peneliti ingin mengetahui pendapat/anggapan masyarakat terkait kasus tersebut berdampak

menjadi komentar negatif, netral atau tetap positif terhadap LPDP itu sendiri. Peneliti dengan mempertimbangkan penelitian terdahulu melakukan penelitian dengan mengkomparasi hasil dari dua metode yakni *Support Vector Machine* (SVM) dan *Naïve Bayes* dalam pengklasifikasian sentimen twitter tersebut. Sedangkan untuk perhitungan akurasi akan dilakukan menggunakan metode *Confusion Matrix* selain itu memvisualisasikan kata-kata yang sering muncul pada bagian sentimen yang dinyatakan positif, negatif maupun netral.

## 2. Metodologi Penelitian

### 2.1. Alur Penelitian

Dalam penelitian ini, kami mengambil tema “LPDP” dalam proses *crawling* Twitter untuk melakukan analisis sentimen yang terdapat di sosial media twitter. Tahap yang akan kami lakukan dalam penelitian ini, ditunjukkan pada Gambar 1:



Gambar 1. Alur Penelitian

Dalam proses penelitian ini kami menggunakan dua *Library Vader Lexicon* yakni pertama berupa bawaan dari NLTK dan kedua merupakan *Library* dari *VaderSentiment*. Pada bagian hasil dan pembahasan akan dijabarkan nilai *compound* yang diperoleh dari kedua *Library Vader Lexicon* tersebut. Data ini menjadi data train/data latih/data yang sudah siap untuk dilakukan pengklasifikasian terhadap dua metode yakni *Naïve Bayes* dan *Support Vector Machine* (SVM).

### 2.2. Twitter

Twitter adalah salah satunya jejaring sosial terbaik dan tetap populer sampai kini. Twitter mempunyai sekitar 330 juta pengguna perbulan yang aktif. Twitter merupakan penyedia platform yang baik dalam berbagi pendapat maupun informasi mengenai topik yang sedang tren dalam perbincangan. Selain itu, pengguna twitter dapat memposting tweetnya secara terbuka. Ini menyediakan platform yang bagus untuk berbagi pendapat dan pandangan pengguna tentang topik yang sedang tren. Twitter tidak memiliki grup lokal, setiap pengguna dapat memposting tweet dan memiliki akses terbuka. Twitter menyediakan lingkungan yang baik untuk analisis sentimen [6].

### 2.3. Analisis Sentimen

Analisis sentimen (SA), atau penggalian opini, didefinisikan sebagai proses menggunakan mesin learning

(ML) dan *natural language processing* (NLP) untuk mengklasifikasikan emosi dalam informasi subjektif. Ini dianggap sebagai salah satu bidang penelitian paling populer di bidang alam pemrosesan bahasa, karena menyediakan sarana untuk meninjau dan menganalisis pendapat yang dipegang oleh sejumlah individu [4]. Untuk mengumpulkan informasi yang terkait dengan jangkauan yang luas populasi, platform media sosial, seperti Twitter adalah sumber berharga dari informasi. Platform ini telah menjadi sangat populer bagi individu untuk bebas mengekspresikan perasaan dan berbagi informasi setiap hari terhadap peristiwa apapun.

### 2.4. Crawling Data Twitter

*Crawling* merupakan proses pengambilan informasi berupa teks yang terdapat dalam sosial media. Proses pengambilan data twitter ini harus mendapatkan akses terlebih dahulu melalui laman <https://developer.twitter.com/en>, tentunya syarat utamanya harus memiliki akun twitter. Pada laman tersebut, nantinya akan mendapatkan *API Key* dan *Token Access* totalnya ada empat [7]. Selain itu, kami memanfaatkan *Library Twitter* dalam proses *crawling* data twitter di Python yakni *Library Tweepy*.

### 2.5. Data Pre-processing

#### a) Case Folding

Mengonversi/mengubah huruf kapital dalam kalimat ke dalam bentuk huruf kecil seluruhnya, ini bertujuan supaya kalimat tersebut memiliki bentuk standar yang sama [8].

#### b) Normalisasi

Menghapus fitur-fitur yang tidak diperlukan dalam proses analisis nantinya, selain itu menggantikan kata yang tidak sesuai EYD menjadi kata yang lebih struktural [8]. Misal kata “sdh” menjadi “sudah”.

#### c) Stopword

Stopword merupakan tahapan penghapusan kata yang terdapat dalam daftar *stoplist* [3]/ atau kata yang tidak memiliki arti penting. Seperti: dan, yang, dengan, akan, dan lain-lain.

#### d) Stemming

*Stemming* merupakan penguraian kata dari berbagai bentukannya untuk menjadi bentuk kata dasar. Maksud lainnya yakni menghilangkan imbuhan-imbuhan yang menempel pada kata dasar tersebut [9].

### 2.6. Language Translation

Tahap ini kami memanfaatkan *Library* dari *Google Translate* untuk melakukan *translate* secara otomatis [6]. Karena kami *crawling* twitter yang Indonesia, maka teks tersebut ditranslate ke Bahasa Inggris. Bertujuan untuk mempermudah dalam tahap proses pelabelan yang akan dilakukan menggunakan *Vader Lexicon*.

### 2.7. Pelabelan Menggunakan Vader Lexicon

Tahap ini kami memanfaatkan Library dari *Google Translate* untuk melakukan translate secara otomatis [6]. Karena kami *crawling* twitter yang Indonesia, maka teks tersebut ditranslate ke Bahasa Inggris. Bertujuan untuk mempermudah dalam tahap proses pelabelan yang akan dilakukan menggunakan *Vader Lexicon*.

2.8. Metode *Naïve Bayes*

*Naïve Bayes* adalah yang paling populer algoritme klasifikasi karena kesederhanaan dan efektivitasnya. Pengklasifikasi ini bekerja sesuai dengan konsep *Teorema Bayes*. Ini semacam pengklasifikasi modul yang mengikuti gagasan probabilitas untuk tujuan klasifikasi [10]. Rumus perhitungan *Naïve Bayes* ditunjukkan pada Persamaan (1) dan (2):

**Probabilitas Prior**

$$P(H) = \frac{N_j}{N} \tag{1}$$

Dimana:

- $N_j$  jumlah data pada kelas
- $N$  jumlah keseluruhan data

**Probabilitas Posterior**

$$P(H|X) = \frac{P(X|H) \cdot P(H)}{P(X)} \tag{2}$$

Dimana:

- X data kelas yang tidak diketahui
- H hipotesis data X kelas tertentu
- P(H|X) Probabilitas Hipotesis H berdasarkan kondisi X (posterior probabilitas)
- P(H) Probabilitas hipotesis H (prior probabilitas)
- P(X) Probabilitas X
- P(X|H) Probabilitas X berdasarkan kondisi pada hipotesis H

2.9. Metode *Support Vector Machine (SVM)*

Mesin vektor dukungan (SVM) diketahui berfungsi baik dalam analisis sentimen. SVM menyelidiki informasi, mencirikan batas pilihan dan menggunakan komponen untuk perhitungan, yang dilakukan di ruang input. Itu informasi penting disajikan dalam dua pengaturan vektor, masing-masing ukuran m. Pada titik ini, setiap datum (dinyatakan sebagai vektor) diurutkan ke dalam kelas. Selanjutnya, mesin mengidentifikasi batas antara dua kelas yang jauh dari tempat manapun di sampel pelatihan. Keterpisahan mencirikan tepi klasifikasi, memperluas tepi mengurangi ambivalen pilihan. Seperti yang ditunjukkan pada, SVM telah terbukti berkinerja lebih efektif daripada pengklasifikasi *Naïve Bayes* di berbagai masalah klasifikasi teks [11]. Rumus perhitungan Support Vector Machine (SVM) dijabarkan pada **Persamaan (3), (4), (5), (6), (7), (8) dan (9)**:

**Perhitungan Matriks**

$$D_{ij} = y_i y_j (K(\vec{x}_i, \vec{x}_j) + \lambda^2) \tag{3}$$

Dimana :

- $D_{ij}$  Elemen matriks data ke-ij
- $y_i$  Kelas atau label data ke-i
- $y_j$  Kelas atau label data ke-j
- $K(\vec{x}_i, \vec{x}_j)$  Fungsi kernel

$\lambda$  Turunan batas teoritis

**Data ke n**

$$E_i = \sum_{j=1}^n a_i D_{ij} \tag{4}$$

$$\delta a_i = \min\{\max[\gamma(1 - E_i), -a_i], c - a_i\} \tag{5}$$

$$a_i = a_i + \delta a_i \tag{6}$$

Dimana :

- $E_i$  Nilai error data ke-i
- $\gamma$  Tingkat pembelajaran
- $\max_{(i)} D_{ij}$  Nilai maksimum diagonal matriks hessian

**Mencari nilai bias (b)**

$$b = -\frac{1}{2} [w \cdot x^+ + w \cdot x^-] \tag{7}$$

**Perhitungan Keputusan**

$$h(x) = \begin{cases} +1, & \text{if } w \cdot x + b \geq 0 \\ -1, & \text{if } w \cdot x + b < 0 \end{cases}$$

jika hasil perhitungan keputusan  $\geq 0$  maka nilai sign  $h(x) = +1$ , termasuk kelas positif sedangkan apabila hasil perhitungan keputusan  $< 0$  maka nilai sign  $h(x)$  nilai -1, termasuk kelas negatif.

$$h(x) = w \cdot x + b \tag{8}$$

Atau

$$h(x) = \sum_{i=1}^m a_i y_i K(x, x_i) + b \tag{9}$$

2.10. *Confusion Matrix*

*Confusion Matrix* ialah metode yang dipergunakan dalam menemukan akurasi, presisi, recall dan F1-Score [12]. Tabel acuan dari confusion matrix ditunjukkan pada Tabel 1:

Tabel 1. Confusion Matrix

		Nilai Actual	
		True	False
Nilai Prediksi	True	TP (True Positif) Corect Result	FP (False Positif) Unexpected Result
	False	FN (False Negatif) Missing Result	TN (True Negatif) Corect Obsence of Result

a) *Recall*

*Recall (sensitivity)* merupakan nilai yang menunjukkan tingkat keberhasilan dalam mengetahui sebuah informasi yang benar tentang data yang dinyatakan negatif atau positif [13]. Rumus *Recall* diperlihatkan pada Persamaan (10):

$$Recall = \frac{TP}{TP+FN} \tag{10}$$

b) *Presisi*

*Presisi (Precision)* merupakan kesesuaian informasi antara yang diminta user dan yang diperoleh dari sistem. Rumus *Presisi* diperlihatkan pada Persamaan (11):

$$Presisi = \frac{TP}{TP+FP} \times 100\% \tag{11}$$

c) *F1-Score*

*F1-Score* ialah sspembobotan dari nilai presisi dan *recall*. Rumus *F1-Score* diperlihatkan pada Persamaan (12):

$$F1 - Score = 2 X \frac{Precision X Recall}{Precision + Recall} \tag{12}$$

d) Akurasi

Akurasi (*accuracy*) merupakan nilai dari hasil actual dan hasil prediksinya sama [14]. Rumus Akurasi diperlihatkan pada Persamaan (13):

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} X 100\% \tag{13}$$

3. Hasil dan Pembahasan

3.1. Pengumpulan Data

Melalui teknik *crawling*, memanfaatkan *library tweepy* dan akses yang didapatkan dari twitter [15]. Hasil proses *crawling* tersebut, kami memperoleh tweet sebanyak 1000 tweet. *Library* dan *API access* yang dibutuhkan seperti yang diperlihatkan pada Gambar 2:

```
!pip install tweet-preprocessor
import tweepy
API_key = "XXXXXXXXXXXXXXXXXXXXXXXXXXXXX"
API_secret = "XXXXXXXXXXXXXXXXXXXXXXXXXXXXX"
access_token = "XXXXXXXXXXXXXXXXXXXXXXXXXXXXX"
access_secret = "XXXXXXXXXXXXXXXXXXXXXXXXXXXXX"
auth = tweepy.OAuthHandler(API_key, API_secret)
auth.set_access_token(access_token, access_secret)
api = tweepy.API(auth, wait_on_rate_limit=True,
wait_on_rate_limit_notify=False, compression=True)
```

Gambar 2. Library dan API Access Twitter

Data hasil *crawling* diperlihatkan pada Tabel 2 dengan memperlihatkan 3 data teratas dan 3 data terbawah:

Tabel 2. Data Hasil Crawling

Tweet	Label
@almadaraaa @DestaHanantio @LPDP_RI @VeritasArdentur Silahkan	1
meskipun awardee2 yg ngakalin akad lpdp nyebelin, tapi gw khawatir keributan ini akan dipelintir ke arah dihapusnya program beasiswa utk diganti jadi program student loan. akad lpdp bisa diperbaiki, tapi fitnah yang dibawa program student loan jauh lebih dahsyat.	1
Dengan masuknya BUMN di Blok Masela, Bapak Jokowi berharap produksi dan lifting migas nasional bisa meningkat. Park Jeongwoo   Haruto   LPDP   #HOBIPALOOZA https://t.co/q0Cok7gBAJ	1
...	
ayo adik-adik, jadikan menkominfo dan segala guyonannya sbg motivasi untuk kalian studi di luar negeri dan gak balik-balik lagi (tp jangan pake lpdp)	1
@dilafarah @petrichoreunia Ya kalo STAN atau ikatan dinas lainnya, sudah diarahin tugas yang mau dia buat apa setelah selesai pendidikan. Kalo LPDP setaunku ya pokoknya "Lu harus balik! Kerjaan? cari sendiri. Kalo mau riset? Cari sponsor sendiri, Alatnya ada gak? Ya cari sendiri."	1
ayo adik-adik, jadikan menkominfo dan segala guyonannya sbg motivasi untuk kalian studi di luar negeri dan gak balik-balik lagi (tp jangan pake lpdp)	1

3.2. Pre-Processing Teks

a) Case Folding

Tahap ini memperlihatkan proses perubahan huruf kapital menjadi huruf kecil atau yang biasa disebut dengan *Lower Case*. Bukan hanya itu, karakter tanda baca dan angka juga dihapus pada tahap ini. Hasil dari proses ini dicontohkan pada Tabel 3:

Tabel 3. Contoh Hasil Case Folding

Tweet	Label
silahkan	1
meskipun awardee yg ngakalin akad lpdp nyebelin tapi gw khawatir keributan ini akan dipelintir ke arah dihapusnya program beasiswa utk diganti jadi program student loan akad lpdp bisa diperbaiki tapi fitnah yang dibawa program student loan jauh lebih dahsyat	
dengan masuknya bum di blok masela bapak jokowi berharap produksi dan lifting migas nasional bisa meningkat	1
park jeongwoo haruto lpdp hobipalooza	

b) Normalisasi Teks

Proses ini mengubah/mengganti kata gaul, slang, singkatan chat ataupun salah penulisan, dirubah menjadi kata yang baku. Daftar kata yang terdapat dalam file *normalisasi* text seperti yang diperlihatkan pada Tabel 4 (4 teratas dan 4 terbawah):

Tabel 4. Daftar Normalisasi

Kata Asal	Kata Baku
singkat	hasil
abis	habis
accent	tekanan
accept	terima
...	
ywdh	ya sudah
ywdhhlh	ya sudahlah
ywis	ya sudah
rp	rupiah

Hasil dari proses ini dicontohkan pada Tabel 5:

Tabel 5. Contoh Hasil Normalisasi

Tweet	Label
silahkan	1
meskipun awardee yang ngakalin akad lpdp nyebelin tapi saya khawatir keributan ini akan dipelintir ke arah dihapusnya program beasiswa untuk diganti jadi program student loan akad lpdp bisa diperbaiki tapi fitnah yang dibawa program student loan jauh lebih dahsyat	
dengan masuknya bum di blok masela bapak jokowi berharap produksi dan lifting migas nasional bisa meningkat	1
park jeongwoo haruto lpdp hobipalooza	

c) Remove Stopwords

Penghapusan kata yang tidak memiliki arti yang berpengaruh dalam kalimat. Hasil dari proses ini dicontohkan pada Tabel 6:

Tabel 6. Contoh Hasil Remove Stopwords

Tweet	Label
silahkan	1
awardee ngakalin akad lpdp nyebelin khawatir keributan dipelintir arah dihapusnya program beasiswa diganti program student loan akad lpdp diperbaiki fitnah dibawa program student loan dahsyat	1
masuknya bumh blok masela jokowi berharap produksi lifting migas nasional meningkat park jeongwoo haruto lpdp hobipalooza	1

d) *Stemming* Sastrawi

Library *Stemming* dari Sastrawi kami manfaatkan untuk proses pengubahan kata dalam kalimat menjadi bentuk kata dasarnya dari berbagai bentukan katanya. Library yang digunakan ditunjukkan pada Gambar 3:

```
# stemming menggunakan sastrawi
!pip install sastrawi
!pip install PySastrawi
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
factory = StemmerFactory()
stemmer = factory.create_stemmer()
def stemming(text):
    text = stemmer.stem(text)
    return text
data['processed'] = data['processed_normal'].apply(lambda x: stemming(x))
data.head()
```

Gambar 3. Source Code Stemming Sastrawi

Hasil dari proses *stemming* sastrawi ini dicontohkan pada Tabel 7:

Tabel 7. Contoh Hasil Stemming

Tweet	Label
silah	1
meski awardee yang ngakalin akad lpdp nyebelin tapi saya khawatir ribut ini akan pelintir ke arah hapus program beasiswa untuk ganti jadi program student loan akad lpdp bisa baik tapi fitnah yang bawa program student loan jauh lebih dahsyat dengan masuk bumh di blok masela bapak jokowi harap produksi dan lifting migas nasional bisa tingkat park jeongwoo haruto lpdp hobipalooza	1

3.3. *Translate* Teks

Di tahap ini, hasil *pre-processing* yang sudah bersih akan ditranslate secara otomatis dengan memanfaatkan Library milik *Google Translate*, seperti yang ditunjukkan pada Gambar 4:

```
#install google translate
!pip3 install googletrans==3..1.0a0
#import library kebutuhan google translate
import googletrans
from googletrans import Translator
translator = Translator()
```

Gambar 4. Source Code Google Translate

Hasil dari Translate Otomatis ini diperlihatkan pada Tabel 8:

Tabel 8. Contoh Hasil Translate

Tweet Indonesia	Tweet Translate	Label
silah	please	1
meski awardee yang ngakalin akad lpdp nyebelin tapi saya khawatir ribut ini akan pelintir ke arah hapus program beasiswa untuk ganti jadi program student loan akad lpdp bisa baik tapi fitnah yang bawa program student loan jauh lebih dahsyat	even though the awardee who broke the lpdp contract is annoying but i'm afraid this fuss will twist towards abolishing the scholarship program to replace it with a student loan program the lpdp contract can be good but the slander that brings the student loan program is far more powerful	1
dengan masuk bumh di blok masela bapak jokowi harap produksi dan lifting migas nasional bisa tingkat park jeongwoo haruto lpdp hobipalooza	by entering the state-owned company in the Masela block, Mr. Jokowi hopes that the national oil and gas production and lifting can be at the level of Park Jeongwoo Haruto LPDP Hobby Palooza	1

3.4. *Vader Lexicon*

Pelabelan kalimat positif, negatif maupun netral sebenarnya juga bisa dilakukan secara manual, namun dalam pengerjaanya harus dilakukan oleh ahli analisis sentiment meliputi:

- 1) Ahli Analisis Sentiment
- 2) Ahli Bahasa atau Longuistik
- 3) Ahli Sosial Media
- 4) Ahli Domain Terkait

Pemilihan ahli yang tepat tergantung pada jenis dan kompleksitas analisis yang dilakukan. Untuk memastikan verifikasi sentimen yang akurat dan menyeluruh, penting untuk melibatkan individu atau tim yang memiliki pengetahuan dan keahlian yang relevan.

Analisis sentimen juga dapat dilakukan tanpa mengikutsertakan para ahli karena perkembangan teknologi dan tersedianya metode pelabelan yang otomatis. Sehingga dapat mengetahui metode pelabelan mana yang lebih cocok dalam penelitian. Maka dari itu dalam penelitian ini yang melibatkan dataset sebanyak 1000 tweet dilabelkan dengan memanfaatkan Library *Vader Lexicon* untuk melakukan pelabelan otomatis. Pelabelan tersebut terbagi menjadi 3 bagian yakni positif, negatif dan netral. Di sini kami menggunakan 2 macam *Vader Lexicon*, yang pertama yakni bawaan dari *NLTK* sedangkan satunya merupakan dari *VaderSentiment* itu sendiri.

1) Vader Lexicon dari NLTK

Gambar 5 merupakan Library yang dibutuhkan dalam *Vader Lexicon* milik NLTK:

```
import nltk
nltk.download('vader_lexicon')
from nltk.sentiment.vader import SentimentIntensityAnalyzer
sid = SentimentIntensityAnalyzer()
```

Gambar 5. Source Code Vader Lexicon dari NLTK  
Penentuan nilai Compound Score tertera pada Gambar 6:

```
df['comp_score'] = df['compound'].apply(lambda c: 'Positif' if c >=0.1
                                         else ('Negatif' if c <0 else 'Netral'))
```

Gambar 6. Source Code Ambang Batas Vader NLTK  
Hasil proses *Vader Lexicon* dari NLTK seperti yang diperlihatkan pada Tabel 9:

Tabel 9. Contoh Hasil Vader Lexicon dari NLTK

text	neg	neu	pos	Compound	label
please	0	0	1	0,3182	Positif
even though the awardee who broke the lpdp contract is annoying but i'm afraid this fuss will twist towards abolishing the scholarship program to replace it with a student loan program the lpdp contract can be good but the slander that brings the student loan program is far more powerful by entering the state-owned company in the Masela block, Mr. Jokowi hopes that the national oil and gas production and lifting can be at the level of Park Jeongwoo Haruto LPDP Hobby Palooza	0,113	0,702	0,185	0,9982	Positif
	0,079	0,845	0,076	-0,0772	Negatif

2) Vader Lexicon dari VaderSentiment

Gambar 7 merupakan Library yang diperlukan untuk memanggil *VaderSentiment*:

```
# lexicon based
!pip install VaderSentiment
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
analyser = SentimentIntensityAnalyzer()
```

Gambar 7. Source Code Vader Lexicon dari VaderSentiment  
Penentuan nilai *Compound Score* tertera pada Gambar 8:

```
#Compound Score lexicon based
tweet_df.loc[tweet_df['Compound_Score'] < 0, 'Sentiments'] = 'Negatif'
tweet_df.loc[tweet_df['Compound_Score'] == 0, 'Sentiments'] = 'Netral'
tweet_df.loc[tweet_df['Compound_Score'] > 0, 'Sentiments'] = 'Positif'
```

Gambar 8. Source Code Ambang Batas VaderSentiment  
Hasil proses *Vader Lexicon* dari NLTK seperti yang diperlihatkan pada Tabel 10:

Tabel 10. Contoh Hasil Vader Lexicon dari VaderSentiment

text	Compound	label
please	0,3182	Positif
even though the awardee who broke the lpdp contract is annoying but i'm afraid this fuss will twist towards abolishing the scholarship program to replace it with a student loan program the lpdp contract can be good but the slander that brings the student loan program is far more powerful by entering the state-owned company in the Masela block, Mr. Jokowi hopes that the national oil and gas production and lifting can be at the level of Park Jeongwoo Haruto LPDP Hobby Palooza	0,7986	Positif
	-0,0258	Negatif

Perbandingan hasil *labeling* otomatis dari kedua *Vader Lexicon* tersebut seperti yang diperlihatkan pada Tabel 11:

Tabel 11. Contoh Hasil Perbandingan dari NLTK dan VaderSentiment

text	NLTK		Vader -Sentiment	
	Compound	label	Compound	label
please	0,3182	Positif	0,3182	Positif
even though the awardee who broke the lpdp contract is annoying but i'm afraid this fuss will twist towards abolishing the scholarship program to replace it with a student loan program the lpdp contract can be good but the slander that brings the student loan program is far more powerful by entering the state-owned company in the Masela block, Mr. Jokowi hopes that the national oil and gas production and lifting can be at the level of Park Jeongwoo Haruto LPDP Hobby Palooza	0,9982	Positif	0,7986	Positif
	-0,0772	Negatif	-0,0258	Negatif

Jeongwoo Haruto LPDP  
Hobby Palooza

### 3.5. Split Data

Sebelum masuk ke proses klasifikasi menggunakan metode. Data akan displit/dibagi dengan perbandingan menjadi 70:30 , 80:20 dan 90:10 (data latih : data tes). Gambar 9 merupakan code untuk melakukan split data:

```
x=data["text"]
y=data["comp_score"]
X_train, X_test, y_train, y_test =
train_test_split(x, y, test_size=0.2, random_state=13,
                shuffle=True , stratify=y)
print(y.value_counts())
```

Gambar 9. Source Code Split Data

### 3.6. Klasifikasi dan Akurasi

Tahap ini, data yang sudah displit/dibagi akan diproses menggunakan dua metode yakni Naïve bayes dan Support Vector Machine (SVM). Hasil perhitungan *Confusion Matrix* diperlihatkan pada Tabel 12 dan Tabel 13:

1) *Vader Lexicon* dari NLTK

Tabel 12. Hasil Confusion Matrix A

Hasil Matrix	Confusion	Naïve Bayes			Support Machine (SVM)		
		70 : 30	80 : 20	90 : 10	70 : 30	80 : 20	90 : 10
Presisi	Neg	0.62	0.64	0.57	0.59	0.71	0.43
	Net	0.17	0.17	1.00	1.00	1.00	1.00
	Pos	0.56	0.62	0.58	0.58	0.66	0.75
Recall	Neg	0.34	0.37	0.40	0.48	0.56	0.60
	Net	0.11	0.08	0.17	0.10	0.25	0.17
	Pos	0.78	0.88	0.88	0.88	0.91	0.86
F1-Score	Neg	0.44	0.47	0.47	0.53	0.63	0.50
	Net	0.13	0.11	0.29	0.17	0.40	0.29
	Pos	0.65	0.73	0.70	0.70	0.77	0.80
Akurasi		0.52	0.58	0.59	0.59	0.69	0.69

2) *Vader Lexicon* dari *VaderSentiment*

Tabel 13. Hasil Confusion Matrix B

Hasil Matrix	Confusion	Naïve Bayes			Support Machine (SVM)		
		70 : 30	80 : 20	90 : 10	70 : 30	80 : 20	90 : 10
Presisi	Neg	0.72	0.71	0.68	0.98	0.94	1.00
	Net	0.44	0.44	0.44	0.72	0.67	1.00
	Pos	0.91	0.89	0.92	0.85	0.86	0.85
Recall	Neg	0.82	0.81	0.84	0.79	0.78	0.84
	Net	0.52	0.50	0.57	0.57	0.62	0.62
	Pos	0.82	0.82	0.79	0.97	0.95	1.00
F1-Score	Neg	0.76	0.76	0.75	0.87	0.86	0.91
	Net	0.48	0.47	0.50	0.63	0.65	0.76
	Pos	0.86	0.86	0.85	0.91	0.90	0.92
Akurasi		0.80	0.80	0.79	0.88	0.87	0.90

### 3) Perbandingan Hasil Klasifikasi

Tabel 14. Hasil Klasifikasi dari Dua Metode

Hasil Klasifikasi	Naïve Bayes			Support Machine (SVM)			
	70 : 30	80 : 20	90 : 10	70 : 30	80 : 20	90 : 10	
NLTK	Neg	29	19	10	27	18	5
	Net	18	12	6	21	12	6
	Pos	49	33	16	48	34	21
Vader Sentiment	Neg	93	62	31	101	65	31
	Net	21	14	7	23	16	13
	Pos	186	124	62	176	199	56

### 3.7. Visualisasi

Tahap ini menampilkan kata-kata yang sering muncul atau dominan muncul pada kalimat tweet dari hasil *crawling* twitter, tentunya sudah melalui tahap *pre-processing*. Ditunjukkan oleh Gambar 10, 11, 12 dan 13:

1) *Vader Lexicon* dari NLTK



Gambar 10. Visualisasi Words Keseluruhan (NLTK)



Gambar 11. Visualisasi Words Positif (NLTK)

2) *Vader Lexicon* dari *VaderSentiment*

Gambar 12. Visualisasi Words Keseluruhan (VaderSentiment)



Gambar 13. Visualisasi Words Positif (VaderSentiment)

Kami menemukan bahwa isi tweet lebih banyak menggunakan bahasa slang/gaul/tidak baku sehingga bahasa-bahasa tersebut masih belum tercover dalam proses normalisasi apalagi bahasa-bahasa tersebut semakin berkembang. Selain itu, isi tweet terkadang tidak sinkron/tidak sama dengan tweet yang mereka retweet, misal iklan namun memakai hashtag LPDP atau meretweet berita/komen tentang LPDP. Ini masih salah satu hal yang membuat hasil analisis sentimen tidak dapat bekerja secara maksimal. Selain itu, pelabelan yang dihasilkan oleh Vader Lexicon dari NLTK dan VaderSentiment memiliki perbedaan dalam penentuan label positif, negatif dan netral pada poin compound score seperti yang sudah dipaparkan pada bagian Hasil dan Pembahasan. Padahal penentuan ambang batas nilai Compound Score untuk label positif, negatif dan netral sudah sama. Ini yang perlu diteliti untuk penelitian selanjutnya, tentang penyebab ketidaksesuaian ini.

## 4. Kesimpulan

Pada kasus sentimen bertema ‐LPDP‐ yang kami teliti menunjukkan bahwa metode pengklasifikasi sentimen yang menggunakan Library Vader Lexicon bawaan NLTK, metode Naive Bayes, dan SVM tidak mencapai tingkat akurasi 70%. Sebaliknya, metode SVM yang menggunakan Library VaderSentiment dari Vader Sentiment mencapai tingkat akurasi 90%, dengan perbandingan 90:10 (data latih : data tes).

Oleh karena itu, metode SVM yang menggunakan Library VaderSentiment dari Vader Sentiment menunjukkan hasil yang lebih baik dalam mengklasifikasi sentimen tweet terkait LPDP, dengan tingkat akurasi yang lebih tinggi daripada metode lain yang diuji.

## Referensi

- [1] R. Sulbakti, *Jamu LPDP: Jurus Ampuh Dapet Beasiswa LPDP*. 2013.
- [2] H. A. Santoso, E. H. Rachmawanto, A. Nugraha, A. A. Nugroho, D. R. I. M. Setiadi, and R. S. Basuki, ‐Hoax classification and sentiment analysis of Indonesian news using Naive Bayes optimization,‐ *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 18, no. 2, pp. 799–806, 2020, doi: 10.12928/TELKOMNIKA.V18I2.14744.
- [3] P. Arsi and R. Waluyo, ‐Analisis Sentimen Wacana Pemindahan Ibu Kota Indonesia Menggunakan Algoritme Support Vector Machine (SVM),‐ *J. Teknol. Inf. dan Ilmu Komput.*, vol. 8, no. 1, p. 147, 2021, doi: 10.25126/jtiik.0813944.
- [4] S. S. Aljameel *et al.*, ‐A sentiment analysis approach to predict an individual’s awareness of the precautionary procedures to prevent covid-19 outbreaks in Saudi Arabia,‐ *Int. J. Environ. Res. Public Health*, vol. 18, no. 1, pp. 1–12, 2021, doi: 10.3390/ijerph18010218.
- [5] A. Rifa’i, H. Sujaini, and D. Prawira, ‐Sentiment Analysis Objek Wisata Kalimantan Barat Pada Google Maps Menggunakan Metode Naive Bayes,‐ *J. Edukasi dan Penelit. Inform.*, vol. 7, no. 3, p. 400, 2021, doi: 10.26418/jp.v7i3.48132.
- [6] K. Arun and A. Srinagesh, ‐Multi-lingual Twitter sentiment analysis using machine learning,‐ *Int. J. Electr. Comput. Eng.*, vol. 10, no. 6, pp. 5992–6000, 2020, doi: 10.11591/ijece.v10i6.pp5992-6000.
- [7] A. K. Fauziyyah, ‐Analisis Sentimen Pandemi Covid19 Pada Streaming Twitter Dengan Text Mining Python,‐ *J. Ilm. SINUS*, vol. 18, no. 2, p. 31, 2020, doi: 10.30646/sinus.v18i2.491.
- [8] N. M. A. J. Astari, Dewa Gede Hendra Divayana, and Gede Indrawan, ‐Analisis Sentimen Dokumen Twitter Mengenai Dampak Virus Corona Menggunakan Metode Naive Bayes Classifier,‐ *J. Sist. dan Inform.*, vol. 15, no. 1, pp. 27–29, 2020, doi: 10.30864/jsi.v15i1.332.
- [9] A. Anggrawan and A. -, ‐Aplikasi Deteksi Kemiripan Tugas Paper,‐ *J. Matrik*, vol. 15, no. 2, p. 5, 2017, doi: 10.30812/matrik.v15i2.39.
- [10] H. Raza, M. Faizan, A. Hamza, A. Mushtaq, and N. Akhtar, ‐Scientific text sentiment analysis using machine learning techniques,‐ *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 12, pp. 157–165, 2019, doi: 10.14569/ijacsa.2019.0101222.
- [11] A. Alsaedi and M. Z. Khan, ‐A study on sentiment analysis



- techniques of Twitter data,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 2, pp. 361–374, 2019, doi: 10.14569/ijacsa.2019.0100248.
- [12] P. Mayadewi and E. Rosely, “Prediksi Nilai Proyek Akhir Mahasiswa Menggunakan Algoritme Klasifikasi Data Mining,” *Semin. Nas. Sist. Inf. Indones.*, no. November, pp. 329–334, 2015.
- [13] M. Azhari, Z. Situmorang, and R. Rosnelly, “Perbandingan Akurasi, Recall, dan Presisi Klasifikasi pada Algoritme C4.5, Random Forest, SVM dan Naive Bayes,” *J. Media Inform. Budidarma*, vol. 5, no. 2, p. 640, 2021, doi: 10.30865/mib.v5i2.2937.
- [14] S. Khomsah and Agus Sasmito Aribowo, “Text-Preprocessing Model Youtube Comments in Indonesian,” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 4, pp. 648–654, 2020, doi: 10.29207/resti.v4i4.2035.
- [15] A. Ristya, C. Chien, and A. Achmad, “Social media sentiment analysis to monitor the performance of vaccination coverage during the early phase of the national COVID-19 vaccine rollout,” *Comput. Methods Programs Biomed.*, vol. 221, p. 106838, 2022, doi: 10.1016/j.cmpb.2022.106838.