Penerapan Metode Naïve Bayes dan Cosine Similarity Dalam Analisis Sentimen Terhadap Platform Film Ilegal di Media Sosial X (Twiter)

Resa Nur Rahmawaty[1] , Didik Indrayana[2],Agung Pambudi[3]
[1,2,3] Universitas Muhammadiyah Sukabumi, Kota Sukabumi, Jawa Barat
resanurrahmawaty15@gmail.com[1], didik.ind@ummi.ac.id[2], agungpambd@ummi.ac.id[3]

Abstrak

Indonesia termasuk salah satu negara dengan pengguna platform streaming ilegal terbanyak di dunia, menurut survei Asia Video *Industry Association's, Coallition Againts Piracy* (CAP) mengungkap bahwa 63% warga Indonesia yang menggunakan streaming online, lebih suka menonton dari platform streaming ilegal secara gratis dengan berbagai macam, konsekuensi disisi lain tindakan menonton secara ilegal termasuk tindakan melanggar hukum karena tidak memiliki izin siar, dimana hal ini dapat menimbulkan berbagai macam jenis komentar ataupun tanggapan dari netizen terhadap maraknya platform streaming ilegal ini, komentar ataupun tanggapan dari netizen dapat berjenis sentimen positif dan negatif. Komentar berupa sentimen ini dapat muncul dari berbagai media, salah satunya media sosial twitter, dimana media ini merupakan salah satu tempat yang berguna untuk menyerukan pendapat,tanggapan ataupun reaksi sentimen. enelitian ini dilakukan untuk mengetahui komentar, tanggapan netizen terhadap platform situs *streaming* ilegal yang akan diklasifikasikan kedalam sentimen positif dan negatif dengan Algoritma *Naïve Bayes Classification* dengan bantuan metode *Cosine Similiarity.* Berdasarkan presentase *sentiment* atau tanggapan pengguna dimedia social x terhadap platform film illegal cenderung negative. Hal tersebut dapat dilihat dari hasilklasifikasi *naïve bayes classifier* dengan *sentiment* negatif sebesar 82,6%. Serta tingkat akurasi pengujian dengan menggunakan confusion matrix menghasilkan akurasi sebesar 68% positif dan 100% negatif. Sementara itu hasil klasifikasi menggunakan metode *cosine similarity* berdasarkan kategori yang paling sering dibicarakan untuk platform film illegal sebesar 65,4% , kategori harga sebesar 5,2%, kategori aksesibilitas sebesar 9,5%, kategori legalitas sebesar 19% dan keamanan sebesar 0,9%.

Kata Kunci: *Naïve Bayes Clasification*, *Cosine Similiarity*, Twitter (X), Platfom Situs *Streaming* Ilegal

PENDAHULUAN

Museum adalah Sejalan dengan perkembangan teknologi informasi yang makin meningkat serta akses internet yang semakin mudah dan biaya aksesnya yang murah, perkembangan dan penggunaan media sosial pun secara global yang terus meningkat dari tahun ke tahunnya. Berdasarkan situs resmi kominfo, sebanyak 63 juta jiwa penduduk Indonesia merupakan pengguna internet dan dari angka tersebut 95 persen menggunakan internet untuk mengakses media sosial.

Media sosial di internet memungkinkan pengguna untuk mengekspresikan diri, terlibat dalam suatu percakapan, bertukar informasi atau bahkan berkomunikasi dengan pengguna lain secara real time. Di Indonesia, twitter (X) menjadi salah satu media sosial yang sangat disukai dengan lebih dari 19,5 juta pengguna dan menduduki peringkat ke 5 sebagai pengguna teratas di seluruh dunia (Kominfo, 2022).

Menurut (Verawati & Audit, 2022) Twitter (X) adalah media sosial microblog yang memungkinkan penggunanya berkomunikasi dalam pesan singkat dengan maksimal 280 karakter. Salah satu fitur di twitter (X) yaitu trending, yang dimana jika suatu isu atau berita yang cukup menarik perhatian pengguna twitter (X) untuk dibagikan, kemungkinan besar isu atau berita tersebut akan trending dan menjadi

topik hangat yang akan dibahas beberapa hari kedepannya.

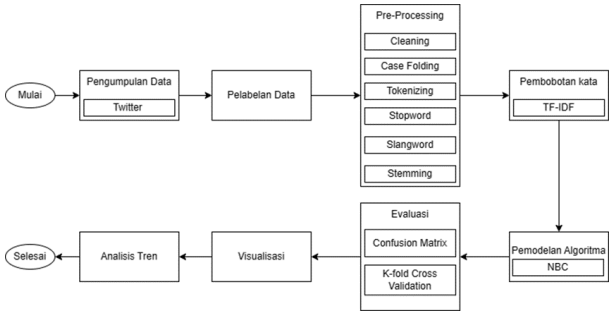
Kecenderungan interaksi yang mudah pada twitter (X) menjadikannya salah satu media sosial yang akrab dengan masyarakat Indonesia. Ditambah dengan fitur yang lebih komprehensif membuat twitter (X) banyak dimanfaatkan masyarakat baik oleh pribadi maupun juga oleh institusi-institusi lain yang memanfaatkan twitter (X) sebagai riset pasar mereka. Salah satu hal yang banyak dibahas saat ini ialah mengenai platform film illegal, mulai dari menonton film di situs yang tidak resmi seperti aplikasi yang menyediakan tontonan film-film atau drama series bajakan. Melihat mulai banyaknya masyarakat baik dari kalangan muda atau tua yang senang menonton film, drama atau series (Mursid & Hartanto, 2019).

Dalam era digital saat ini, media sosial telah menjadi platform yang penting bagi pengguna internet untuk berbagi pendapat, informasi, dan pengalaman mereka. Salah satu topik yang sering dibahas di media sosial adalah film. Namun, ada beberapa platform film ilegal yang menawarkan film-film tanpa izin dari pemegang hak cipta. Keberadaan platform film ilegal ini menjadi perhatian karena dapat merugikan industri film dan pemegang hak cipta, tahun 2019 pemerintah Indonesia tercatat telah berhasil memblokir sebanyak 1.946 platform streaming illegal, namun karena banyaknya peminat di Indonesia eksistensi platform streaming illegal ini menjadi tidak mudah dimusnahkan. Dengan mengunggulkan fasilitas utamanya yaitu memberikan akses secara gratis kepada pengguna, tentunya hal ini memberikan pertimbangan bagi netizen dan menimbulkan berbagai macam komentar, tanggapan serta reaksi sentimen terhadap eksistensi platform streaming illegal tersebut (Wibowo et al., 2023).

Agar mendapat sebuah gambaran mengenai pemecahan masalah ini maka dilakukan analisis terhadap penelitian terdahulu yang memiliki permasalahan sejenis sehingga dapat digunakan sebagai tolak ukur peneliti untuk melakukan penelitian ini. Seperti pada jurnal yang ditulis oleh Widiyanto Tri Handoko dan kawan-kawan yang berjudul "Klasifikasi Opini Pengguna Media Sosial Twitter Terhadap JNT di Indonesia dengan Algoritma Decision Tree" menghasilkan nilai akurasi sebesar 94,12 persen dengan rasio perbandingan untuk data training dan data testing sebesar 90:10 (Handoko et al., 2022).

METODE

Tahapan penelitian yang dilakukan ini berdasarkan metode text mining dan juga menggunakan algoritma Naïve Bayes Classifier yang dimulai dari pengumpulan data, pelabelan data, preprocessing, pembobotan kata, pemodelan algoritma, evaluasi, visualisasi dan analisis tren



Gambar 1. Metodologi Penelitian

A. Pengumpulan Data

Data yang dipakai dalam penelitian ini berasal dari twitter (X). Pengambilan data dilakukan dengan melakukan *crawling* menggunakan bahasa pemrograman phyton dari twitter API mengenai beberapa platform *streaming* film. Kemudian data yang telah terkumpul, disimpan sesuai platform. Kemudian data dibagi menjadi data latih 80% dan data uji 20% pada masing-masing dataset.

1. Pelabelan Data

   Di tahapan ini, data training diberi label dengan setiap teks dokumen diberi label pro atau kontra. Proses pelabelan ini dilakukan secara manual.
2. *Pre-Processing*

Tahapan berikutnya adalah proses *pre-processing* data yang merupakan tahapan yang penting dilakukan. Proses ini dilakukan dengan tujuan untuk mengurangi atribut yang tidak diperlukan dalam proses klasifikasi nantinya. Proses *pre-processing* terdiri dari proses *cleaning, case folding, stopword, slangword, tokenizing, dan stemming* terhadap data mentah yang masih kotor.

3. Pembobotan Kata

   Pada tahapan ini memiliki tujuan untuk meminimalkan informasi yang tidak relevan dan memaksimalkan informasi yang relevan dalam membedakan antar kelas agar memaksimalkan hasil akurasi klasifikasinya. Di tahapan ini menggunakan *Term Frequency-Inverse Document Frequency yang* dimana untuk menghitung frekuensi setiap *term* dalam sebuah dokumen dan frekuensi term tersebut di seluruh koleksi dokumen.

4. Pemodelan Algoritma

   Langkah ini dilakukan untuk proses perhitungan klasifikasi menggunakan metode algoritma *naïve bayes classifier* dengan bahasa pemrograman *phyton*.

5. Evaluasi

   Tahap selanjutnya adalah evaluasi dimana langkah ini dilakukan untuk mengetahui akurasi performa dari algoritma naïve bayes menggunakan confusion matrix serta *K-fold Cross Validation.*

6. Visualisasi

   Tahapan selanjutnya adalah visualisasi. Dimana tahap ini bertujuan untuk memudahkan pemahaman sehingga data dipresentasikan dalam bentuk grafik yang berfungsi untuk mengekstrak informasi dari objek yang paling sering muncul. Hal ini memungkinkan untuk menyaring tweet dalam dataset dan menampilkan informasi yang dibutuhkan dengan lebih mudah.

7. Analisis Tren

   Terakhir adalah tahapan analisis tren. Tahapan ini dilakukan untuk mengetahui pola perubahan yang terjadi pada sentimen terhadap platform film ilegal pada beberapa bulan ini.

HASIL DAN PEMBAHASAN

A. Pengumpulan Data

Data yang digunakan merupakan data berupa tweet yang berisikan tanggapan penonton terhadap platform film yang diambil dari https://x.com/ dengan kata kunci platform pembajakan film yang sering digunakan. Pengumpulan data dilakukan dengan cara crawling menggunakan Bahasa Pemrograman phyton pada *Google Colab*. Dalam proses *crawling* diperlukan *twitter auth token* yang diperoleh dari akun media sosial twitter pribadi untuk proses autentikasi.

```
#@title Twitter Auth Token

twitter_auth_token = '18ce7edd472f5b1ba57767e3d59ac19cadd3da7b'
```

Gambar 2. *Twitter Auth Token*

Kemudian dilakukan proses *crawling* data pada twitter dengan menggunakan search keyword indoxxi illegal dalam Bahasa Indonesia. Proses *crawling* tersebut dilakukan secara bertahap karena terdapat Batasan dari setiap pengambilan data. Kemudian data tersebut disimpan ke dalam file dengan format csv.

```
# Crawl Data

filename = 'indoxxi.csv'
search_keyword = 'indoxxi ilegal lang:id'
limit = 1500

!npx --yes tweet-harvest@2.2.8 -o "{filename}" -s
"{search_keyword}" -l {limit} --token {twitter_auth_token}
```

Gambar 3. Kode Program *Crawling* Data

Setelah proses crawling dilakukan sejak bulan Desember 2023 sampai dengan Februari 2024, menghasilkan dataset dengan jumlah 1506 data.

B. Data Labeling

Sebelum masuk pada tahap *pre-processing*, hasil *crawling* data yang disimpan kedalam bentuk csv akan diberi label secara manual oleh peneliti dengan

mengkategorikan masing-masing dokumen pada dataset tersebut masuk kedalam kelas positif atau negatif. Serta dilakukan seleksi untuk data yang tidak berkaitan dengan kata kunci yang digunakan.

Tabel 1. Contoh Pelabelan Data

Tweet	Sentimen
Harus dibagun kesadaran di Tengah Masyarakat bahwa menonton film bajakan sama dengan mencuri. https://twitter .com/kompascom/status/	Positif
Nonton film di tele/web kualitas suaranya tidak jelas walau sudah pakai headset https://t.co/XDe6vzaOo4"	Negatif

Berikut adalah tabel yang berisikan jumlah dari setiap data yang berlabel positif dan negatif pada setiap dataset.

Tabel 2. Tabel Jumlah Data

Dataset	Positif	Negatif	Jumlah Keseluruhan
Platfom	800	702	1502

C. Pre-Processing

Terdapat beberapa bagian pada tahap pre-processing data yaitu diawali dengan *cleaning, case folding, tokenizing, filtering,* (*stopwords* dan *slangword*) dan terakhir *stemming.*

1. *Cleaning*

Cleaning merupakan proses awal dari *pre-processing* yang bertujuan untuk menghilangkan teks atau data yang tidak valid serta tanda baca, hastag, *link,* mention, angka dan atribut lainnya yang tidak memiliki arti penting pada dataset.

Tabel 3. *Cleaning* Data

Sebelum *Cleaning*	Setelah *Cleaning*
@leyonlovesyuta Awokwok udh kya web film bajakan lainnya saja ini telegram lengkap bgt https://t.co/Xtr70zaO04	Awokwok udh kya web film bajakan lainnya saja ini telegram lengkap bgt

2. *Case Folding*

Tahapan ini bertujuan untuk mengkonversi huruf kapital pada teks komentar atau tweet menjadi huruf kecil.

Tabel 5. *Case Folding*

Sebelum *Case Folding*	Setelah *Case Folding*
Awokwok udh kya web film bajakan lainnya saja ini telegram lengkap bgt	awokwok udh kya web film bajakan lainnya saja ini telegram lengkap bgt

3. *Tokenizing*

Proses tokenizing dilakukan dengan tujuan untuk membagi teks dengan cara memotong kata-kata yang dipisahkan spasi menjadi beberapa bagian tersendiri. Pada tahapan ini memanfaatkan *Library Natural Language Toolkit* (NLTK) pada Bahasa pemrograman *phyton.*

Tabel 6. *Tokenizing*

Sebelum *Tokenizing*	Setelah *Tokenizing*
awokwok udh kya web film bajakan lainnya saja ini telegram lengkap bgt	‘awokwok’, ‘udh’, ‘kya’, ‘web’, ’film’, ‘bajakan’, ‘ lainnya’, ‘saja’, ‘ini’, ‘telegram’, ‘lengkap’, ‘bgt’

4. *Stopword*

Stopword bertujuan untuk menghilangkan kata-kata yang tidak bermakna *(stoplist)* atau kata-kata yang tidak mempengaruhi keakuratan proses klasifikasi, seperti kata

penghubung dan akan digantikan dengan spasi dan menyimpan kata-kata penting *(wordlist).*

Tabel 7. *Stopword*

Sebelum *Stopword*	**Setelah *Stopword***
‘awokwok’, ‘udh’, ‘kya’, ‘web’, ’film’, ‘bajakan’, ‘ ‘ lainnya’, ‘saja’, ‘ini’, ‘telegram’, ‘lengkap’, ‘bgt’	‘udh’, ‘kya’, ‘web’, ’film’, ‘bajakan’, ‘ lainnya’, ‘saja’, ‘telegram’, ‘lengkap’, ‘bgt’

5. *Slangword*

Proses ini bertujuan untuk mengganti kata-kata *slang* atau informal yang biasa digunakan dalam kehidupan sehari-hari kedalam bentuk yang sebenarnya.

Tabel 8. *Slangword*

Sebelum *Slangword*	**Setelah *Slangword***
pny	punya
kya	kaya
bgt	banget
bnyk	banyak
utk	untuk
kya	kaya

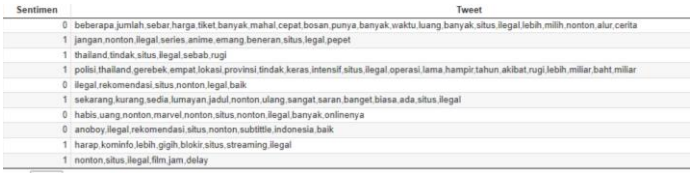
6. *Stemming*

Proses ini bertujuan untuk menghilangkan dan mengembalikannya ke bentuk dasarnya. Dalam proses *stemming* Bahasa Indonesia, dilakukan dengan cara menghilangkan sufiks dan prefix dengan bantuan *library* sastrawi pada *Google Colab.*

Tabel 9. *Stemming*

Sebelum *Stemming*	**Setelah *Stemming***
Sayangnya, saya merasa platform film ini kurang terstruktur dan sulit untuk diikuti	Sayangnya, saya rasa platform film ini kurang struktur dan sulit untuk diikuti

Setelah semua dataset dilakukan proses prepocessing, masing-masing hasilnya akan disimpan dalam bentuk file csv yang dimana hasilnya dapat dilihat sebagai berikut.



Sentimen	Tweet
0	beberapa,jumlah,sebar,harga,tiket,banyak,mahal,cepat,bosan,punya,banyak,waktu,luang,banyak,situs,ilegal,lebih,milih,nonton,alur,cerita
1	jangan,nonton,ilegal,series,anime,emang,beneran,situs,legal,pepet
1	thailand,tindak,situs,ilegal,sebab,rugi
1	polisi,thailand,gerebek,empat,lokasi,provinsi,tindak,keras,intensif,situs,ilegal,operasi,lama,hampir,tahun,akibat,rugi,lebih,miliar,baht,miliar
0	ilegal,rekomendasi,situs,nonton,legal,baik
1	sekarang,kurang,sedia,lumayan,jadul,nonton,ulang,sangat,saran,banget,biasa,ada,situs,ilegal
0	habis,uang,nonton,marvel,nonton,situs,nonton,ilegal,banyak,onlinenya
0	anoboy,ilegal,rekomendasi,situs,nonton,subtitle,indonesia,baik
1	harap,kominfo,lebih,gigih,blokir,situs,streaming,ilegal
1	nonton,situs,ilegal,film,jam,delay

Gambar 4. Data Hasil *Pre-Processing*

D. Pembobotoan Kata

Setelah proses prepocessing selesai, proses selanjutnya adalah menghitung bobot pada suatu kata dalam kalimat dengan menggunakan algoritma TF-IDF atau *frequency inverse document frecuency.* Berikut Sampel dokumen yang dimana merupakan sampel dokumen dari dataset *Platform* film illegal..

Tabel 10. *Sample Data*

Dokumen	**Input**	**Label**
D1	Ada beberapa alas an menurut gw jumlah bioskop dan penyebarannya harga tiket yang buat banyak org mahal cepat bosan tidak punya banyak waktu luang masih banyak yang suka situs illegal orang lebih milih nonton alur cerita film nya saja	Negatif
D2	Jangan nonton illegal kecuali kalua film series anime yang kamu cari emang beneran gak ada di situs legal dan kepepet	Positif
D3	Lk illegal ini dia rekomendasi situs nonton film illegal dan terbaik	Negatif
D4	Gabakal lagi ngabisin duit gw buat nnton film marvel di bioskop gw bakal nnton di situs nnton film illegal yg byk iklan judi onlinenya	Negatif
D5	Gua nonton film di situs illegal filmnya Cuma sejam tapi popup iklan dan delay ampe hari	Positif

Berikut adalah dokumen yang telah melalui tahapan *pre-processing* yang selanjutnya akan dibagi menjadi beberapa token tertentu

Tabel 11. Dokumen Token

Dokumen	Input	Label
D1	Beberapa jumlah penyebarannya harga tiket banyak mahal cepat bosan punya banyak waktu luang banyak situs illegal lebih milih nonton alur cerita	Negatif
D2	Jangan nonton illegal series anime emang beneran situs legal kepepet	Positif
D3	Illegal rekomendasi situs nonton legal terbaik	Negatif
D4	Habis uang nonton marvel situs illegal banyak iklan onlinenya	Negatif
D5	Nonton situs illegal filmnya sejam delay	Positif

1. *Term Frequency* (TF)

Hal pertama yang dilakukan adalah menghitung frekuensi kemunculan setiap kata dalam setiap dokumen. Dari dokumen diatas dapat dilihat perhitungan term frequency seperti pada tabel berikut

Tabel 12.Frekuensi Dokumen

Dokumen	TF					DF
	D1	D2	D3	D4	D5	
Berapa	1	0	0	0	0	1
Jumlah	1	0	0	0	0	1
Penyebarannya	1	0	0	0	0	1
Harga	1	0	0	0	0	1
Tiket	1	0	0	0	0	1
Banyak	3	0	0	1	0	4
Mahal	1	0	0	0	0	1
Cepat	1	0	0	0	0	1
Bosan	1	0	0	0	0	1
Punya	1	0	0	0	0	1
Waktu	1	0	0	0	0	1
Luang	1	0	0	0	0	1
Situs	1	1	1	1	1	5
Illegal	1	1	1	1	1	5
Lebih	1	0	0	0	0	1
Milih	1	0	0	0	0	1
Nonton	1	1	1	1	1	5
Alur	1	0	0	0	0	1
Cerita	1	0	0	0	0	1
Jangan	0	1	0	0	0	1
Series	0	1	0	0	0	1
Anime	0	1	0	0	0	1
Emang	0	1	0	0	0	1
Beneran	0	1	0	0	0	1
Legal	0	0	1	0	0	1
Kepepet	0	1	0	0	0	1
Rekomendasi	0	0	1	0	0	1
Terbaik	0	0	1	0	0	1
Habis	0	0	0	1	0	1
Uang	0	0	0	1	0	1
Marvel	0	0	0	1	0	1
Iklan	0	0	0	1	0	1
Onlinenya	0	0	0	1	0	1
Film	0	0	0	0	1	1
Sejam	0	0	0	0	1	1
Delay	0	0	0	0	1	1

Pada tabel diatas jika suatu dokumen mengandung *term* tertentu, akan diberikan nilai sesuai dengan jumlah term yang muncul pada dokumen tersebut dan jika tidak maka akan diberi nilai 0. Seperti *term* ‘ilegal’, muncul di 5 dokumen yaitu D1, D2, D3, D4 dan D5 dengan kemunculan term pada dokumen-dokumen tersebut sekali sehingga diberi nilai 1.

2. *Inverse Document Frequency* (IDF)

Selanjutnya dilakukan perhitungan IDF pada dokumen dalam tabel di atas dengan jumlah dokumen pada tabel 4.9, dan menghitung jumlah kemunculan setiap kata pada dokumen dengan menggunakan persamaan 2.2 adalah sebagai berikut :

$$IDF\ (banyak) = log\frac{N}{dfilm} = log\frac{5}{1} = 0{,}70$$

Tabel 13. Hasil Perhitungan IDF

Term	DF	IDFLog(n/df
Berapa	1	0,70
Jumlah	1	0,70
Penyebarannya	1	0,70
Harga	1	0,70
Tiket	1	0,70
Banyak	4	0,09
Mahal	1	0,70
Cepat	1	0,70
Bosan	1	0,70

Punya	1	0,70
Waktu	1	0,70
Luang	1	0,70
Situs	5	0
Illegal	5	0
Lebih	1	0,70
Milih	1	0,70
Nonton	5	0
Alur	1	0,70
Cerita	1	0,70
Jangan	1	0,70
Series	1	0,70
Anime	1	0,70
Emang	1	0,70
Beneran	1	0,70
Legal	1	0,70
Kepepet	1	0,70
Rekomendasi	1	0,70
Terbaik	1	0,70
Habis	1	0,70
Uang	1	0,70
Marvel	1	0,70
Iklan	1	0,70
Onlinenya	1	0,70
Film	1	0,70
Sejam	1	0,70
Delay	1	0,40

3. TF-IDF

Setelah nilai TF dan IDF nya ditemukan, selanjutnya adalah menghitung nilai TF-IDF dengan menggunakan persamaan 2.3. pada contoh berikut term yang digunakan adalah term ‘film’ yang akan dihitung nilai TF-IDF nya disetiap dokumen

$$D1 = W_{filmD1} = 0\ x\ 0{,}70 = \ 0$$

$$D2 = W_{filmD2} = 0\ x\ 0{,}70 = \ 0$$

$$D3 = W_{filmD3} = 0\ x\ 0{,}70 = \ 0$$

$$D4 = W_{filmD4} = 0\ x\ 0{,}70 = \ 0$$

$$D5 = W_{filmD5} = 1\ x\ 0{,}70 = \ 0{,}70$$

Gambar 5. Hasil Perhitungan TF-IDF

Tabel 14. Hasil Perhitungan TF-IDF

Term	**IDFLog(n/df**	**TF-IDF**				
		D1	**D2**	**D3**	**D4**	**D5**
Berapa	0,70	0,70	0	0	0	0
Jumlah	0,70	0,70	0	0	0	0
Penyebarannya	0,70	0,70	0	0	0	0
Harga	0,70	0,70	0	0	0	0
Tiket	0,70	0,70	0	0	0	0
Banyak	0,09	0,09	0	0	0,09	0
Mahal	0,70	0,70	0	0	0	0
Cepat	0,70	0,70	0	0	0	0
Bosan	0,70	0,70	0	0	0	0
Punya	0,70	0,70	0	0	0	0
Waktu	0,70	0,70	0	0	0	0
Luang	0,70	0,70	0	0	0	0
Situs	0	0	0	0	0	0
Illegal	0	0	0	0	0	0
Lebih	0,70	0,70	0	0	0	0
Milih	0,70	0,70	0	0	0	0
Nonton	0	0	0	0	0	0
Alur	0,70	0,70	0	0	0	0
Cerita	0,70	0,70	0	0	0	0
Jangan	0,70	0	0,70	0	0	0
Series	0,70	0	0,70	0	0	0
Anime	0,70	0	0,70	0	0	0
Emang	0,70	0	0,70	0	0	0
Beneran	0,70	0	0,70	0	0	0
Legal	0,70	0	0	0,70	0	0
Kepepet	0,70	0	0,70	0	0	0
Rekomendasi	0,70	0	0	0,70	0	0
Terbaik	0,70	0	0	0,70	0	0
Habis	0,70	0	0	0	0,70	0
Uang	0,70	0	0	0	0,70	0
Marvel	0,70	0	0	0	0,70	0
Iklan	0,70	0	0	0	0,70	0
Onlinenya	0,70	0	0	0	0,70	0
Film	0,70	0	0	0	0	0,70
Sejam	0,70	0	0	0	0	0,70
Delay	0,40	0	0	0	0	0,40
Nilai Bobot Setiap Dokumen		**10,59**	**4,2**	**2,1**	**3,59**	**1,8**

E. Pemodelan Algoritma *Naïve Bayes Clasifier*

Pemodelan algoritma adalah tahap yang dilakukan setelah data sesuai dengan model yang diperlukan dan sudah melalui tahapan sebelumnya yaitu pre-processing dan pembobotan kata menggunakan TF-IDF. Selanjutnya dilakukan proses klasifikasi menggunakan naïve bayes classifier. Selama proses klasifikasi ini digunakan beberapa library seperti berikut ini.

```
import pandas as pd
import re
import numpy as np
import joblib
import pickle
from sklearn import model_selection
from sklearn.linear_model import LogisticRegression
from sklearn.multiclass import OneVsRestClassifier
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import classification_report
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.pipeline import Pipeline
from sklearn.metrics import accuracy_score
import csv
from google.colab import drive
drive.mount('/content/gdrive')
```

Gambar 7. Library Pemodelan algoritma

Selanjutnya data-data tersebut akan dibagi menjadi 2 yaitu dimana 80% data akan digunakan sebagai data training dan 20% sisanya akan digunakan sebagai data testing. Data yang telah selesai melalui tahap pembobotan kata menggunakan TF-IDF, Langkah selanjutnya adalah melakukan klasifikasi dengan menggunakan *naïve bayes classifier*. Setelah dilatih, model akan disimpan dalam format pickle di google drive. Kemudian, file *pickle* tersebut akan dimuat Kembali sebagai data model dan akan digunakan untuk melakukan klasifikasi pada data uji. Berikut kode program untuk melakukan klasifikasi dengan menggunakan *naïve bayes classifier*.

```
For train_index, test_index in kf.split(data):
    X = data['Tweet'][train_index]
    Y = data['Sentimen'][train_index]

    x_train, x_test, y_train, y_test = split(X, Y,
test_size=0.2, random_state=42)

    vectorizer = TfidfVectorizer()
    train_vector = vectorizer.fit_transform(x_train)
    test_vector = vectorizer.transform(x_test)

clf = MultinomialNB ()
clf_train = clf.fit(train_vector, y_train)

save_model =
open('/content/gdrive/MyDrive/Data_Skripsi/model_NBC.pickle',
'wb')
pickle.dump(clf_train, save_model)
save_model.close()

clf_model =
open('/content/gdrive/MyDrive/Data_Skripsi/model_NBC.pickle',
'rb')
clfm = pickle.load(clf_model)

y_pred = clfm.predict (test_vector)
```

Gambar 8. Kode Program *Naïve Bayes Clasifier*

F. Evaluasi

1. *Confusion Matrik*

Hasil analisis sentimen menggunakan naïve bayes di evaluasi dengan menghitung presisi, recall dan FI-score. Hal ini dilakukan untuk mengetahui akurasi performa dari algoritma naïve bayes classifier pada setiap dataset. Untuk kode program library confusion matrix dan pengujian algoritma dapat dilihat pada gambar berikut.

```
cf_matrix = confusion_matrix(y_test, y_pred)
print("Accuracy:" , accuracy_score(y_test, y_pred))
print(f'Confusion Matrix:\n {confusion_matrix(y_test,
y_pred)}')
```

Gambar 9.Kode Program *Library Confusion Matrik*

Dari kode program diatas dihasilkan nilai akurasi dan tabel confusion matrix untuk setiap dataset yang dapat dilihat seperti berikut.

```
Accuracy: 0.6620689655172414
Confusion Matrix:
 [[16 47]
 [ 2 80]]
```

Gambar 10. Nilai Akurasi *Confusion Matrik*

```
Accuracy: 0.6620689655172414
Confusion Matrix:
 [[16 47]
 [ 2 80]]
```

Gambar 11. Nilai Kinerja Pemodelan Klasifikasi

	precision	recall	f1-score	support
0	1.00	0.12	0.22	57
1	0.68	1.00	0.81	104
accuracy			0.69	161
macro avg	0.84	0.56	0.51	161
weighted avg	0.79	0.69	0.60	161

Gambar 12. Nilai Kinerja Setiap Label *Dataset*

2. *K-Fold Cross Validarion*

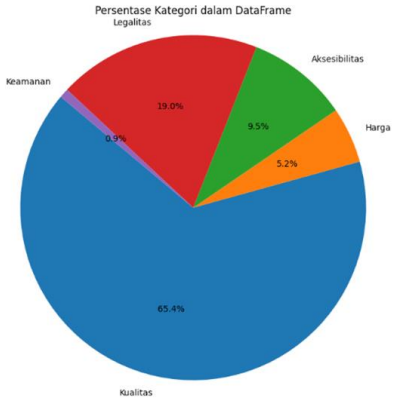
K-fold Cross Validation merupakan tahap evaluasi kedua setelah confusion matrix yang bertujuan untuk memaksimalkan hasil uji dan evaluasi kinerja algoritma, serta menilai tingkat akurasi secara keseluruhan.

```
Iteration       Accuracy        Precision       Recall          F-Measure
-------------------------------------------------------------------------
Iteration-0     0.69444         0.80000         0.61706         0.59259
Iteration-1     0.67586         0.69302         0.57260         0.54387
Iteration-2     0.68276         0.65630         0.58490         0.57505
Iteration-3     0.68276         0.73704         0.56617         0.52816
Iteration-4     0.68966         0.76445         0.59798         0.57210
Iteration-5     0.68276         0.76553         0.55350         0.50431
Iteration-6     0.68276         0.78333         0.57784         0.53886
Iteration-7     0.68276         0.79253         0.60632         0.57505
Iteration-8     0.66207         0.78270         0.61295         0.57376
Iteration-9     0.66207         0.75941         0.61479         0.58031

Average Accuracy  : 0.66207 / 66.20700000000001
Average Precision : 0.75941 / 75.941
Average Recall    : 0.61479 / 61.47899999999999
Average F-Measure : 0.58031 / 58.031

              precision    recall  f1-score   support

           0       0.89      0.25      0.40        63
           1       0.63      0.98      0.77        82

    accuracy                           0.66       145
   macro avg       0.76      0.61      0.58       145
weighted avg       0.74      0.66      0.60       145
```

Gambar 13. *K-Fold Cross Validation* Pada Dataset Platfom

G. Pengklasifikasian Kategori (*Cosine Similiarity)*

Tahap ini bertujuan untuk mengklasifikasikan komentar atau entiment mengenai platform film ke dalam lima jenis kategori yaitu kualitas, harga, aksesibilitas, legalitas dan keamanan. Tahapan ini dilakukan dengan cara membandingkan similaritas antar komentar pelanggan dengan tiga contoh dokumen yang menjadi kata kunci. Lalu jika nilainya mendekati 1 maka dokumen dinyatakan mirip dan jika hasilnya 0 maka dokumen dinyatakan sebaliknya. Berikut adalah kode program untuk mengklasifikasikan komentar menggunakan metode *cosine similarity*.

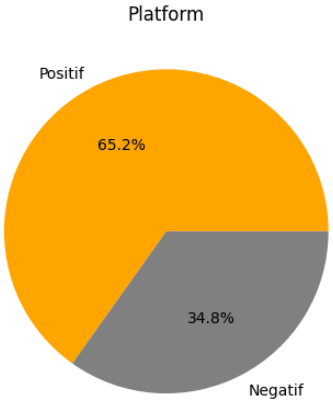


Gambar 14. Presentase Komentar Berdasarkan Kategori

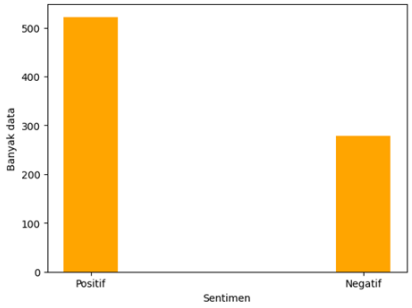
Berdasarkan gambar di atas dapat terlihat perbandingan bahwa presentase komentar pada dataset platform indoxxi dengan kategori kualitas lebih tinggi yaitu 65,4% , kategori harga sebesar 5,2%, kategori aksesibilitas sebesar 9,5%, kategori legalitas sebesar 19% dan keamanan sebesar 0,9%.

1. Visualisasi

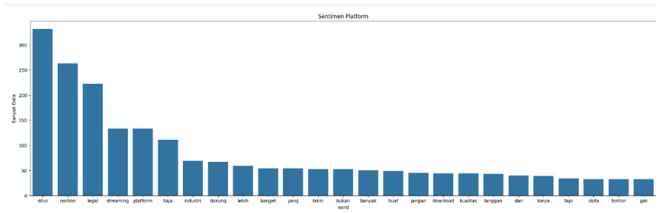
Tahap visualisasi yaitu tahapan yang bertujuan untuk membuat visualisasi hasil akhir klasifikasi kedalam bentuk grafik dengan menggunakan *library matpolib* dan *seaborn* untuk proses visualisasi data. Sehingga dapat membantu dalam menyampaikan informasi yang ingin disampaikan dengan lebih mudah.



Gambar 15. *Pie Chart* Visualisasi Presentase Setiap Label



Gambar 16. Visualisasi Perbandingan Presentase



Gambar 17. Visualisasi *Term* Yang Sering Muncul

Gambar 18. *Wordcloud Sentiment Positif*

Gambar 19. *Worldcloud Sentiment Negatif*

SIMPULAN

Dalam penelitian yang telah penulis lakukan terkait Penerapan Metode Naïve bayes dan Cosine Similarity Dalam Analisis Sentimen Terhadap Platform Film Ilegal di Media Sosial X ini dapat menghasilkan beberapa kesimpulan diantaranya yaitu:

1. Bahwa presentase sentiment atau tanggapan pengguna dimedia social x terhadap platform film illegal cenderung negative. Hal tersebut dapat dilihat dari hasilklasifikasi naïve bayes classifier dengan sentiment negatif sebesar 82,6%. Serta tingkat akurasi pengujian dengan menggunakan confusion matrix menghasilkan akurasi sebesar 68% positif dan 100% negatif.
2. Sementara itu hasil klasifikasi menggunakan metode cosine similarity berdasarkan kategori yang paling sering dibicarakan untuk platform film illegal sebesar 65,4% , kategori harga sebesar 5,2%, kategori aksesibilitas sebesar 9,5%, kategori legalitas sebesar 19% dan keamanan sebesar 0,9%.
3. Penerapan K-fold cross validation pada Penelitian ini menghasilkan nilai akurasi dengan presentase sebesar 63% untuk komentar positif dan 89% komentar negatif

Pada penelitian ini, banyak kerentanan yang masih bisa diperbaiki. Oleh karena itu ada beberapa saran yang diberikan untuk perbaikan pada penelitian berikutnya.

1. Memperbanyak objek platform film yang digunakan sehingga dapat melihat perbandingan sentimen dengan lebih luas.
2. Melakukan pencarian data lebih luas lagi dengan memanfaatkan berbagai media sosial dan platform e-commerce lainnya.

UCAPAN TERIMA KASIH

Terima kasih yang tulus kepada semua pihak yang telah memberikan bimbingan, dukungan, dan motivasi selama penulisan skripsi ini, terutama kepada Bapak/Ibu pembimbing, keluarga tercinta, teman-teman sejawat, serta semua yang turut serta dalam penelitian ini. Segala dedikasi dan doa yang diberikan telah menjadi sumber inspirasi dan kekuatan bagi saya. Semoga hasil penelitian ini dapat bermanfaat bagi perkembangan ilmu pengetahuan di masa depan. Terima kasih atas segala berkat dan rahmat-Nya.

DAFTAR PUSTAKA

Anshari, I. N. (2019). Sirkulasi Film dan Program Televisi di Era Digital: Studi Kasus Praktik Download dan Streaming melalui Situs Bajakan. *Komuniti : Jurnal Komunikasi Dan Teknologi Informasi*, *10*(2), 88–102. https://doi.org/10.23917/komuniti.v10i2.7125

Azhar. (2019). *Analisis Kinerja Algoritma Naïve Bayes dan K-Nearest Neighbor pada Sentimen Analisis dengan pendekatan Lexicon di Media Twitter*. Universitas Islam Negeri Syarif Hidayatullah.

Damuri, A., Riyanto, U., Rusdianto, H., & Aminudin, M. (2021). Implementasi Data Mining dengan Algoritma Naïve Bayes Untuk Klasifikasi Kelayakan Penerima Bantuan Sembako. *Jurnal Riset Komputer*, *8*(6), 219–225. https://doi.org/10.30865/jurikom.v8i6.3655

Darwis, D., Siskawati, N., & Abidin, Z. (2021). Penerapan Algoritma Naive Bayes Untuk Analisis Sentimen Review Data Twitter Bmkg Nasional. *Jurnal Tekno Kompak*, *15*(1), 131. https://doi.org/10.33365/jtk.v15i1.744

Fath, M. K. Al. (2018). *Analisis Sentimen*

Komentar Kebijakan Full Day School dari Facebook Page Kemendikbud RI Menggunakan Algoritma Naïve Bayes Classifier. Universitas Islam Negeri Syarif Hidayatullah.

Firmansyah, Z., & Puspitasari, N. F. (2021). Analisis Sentimen Masyarakat Terhadap Vaksinasi Covid-19 Berdasarkan Opini Pada Twitter Menggunakan Algoritma Naive Bayes. *Jurnal Teknik Informatika*, *14*(2), 171–178. https://doi.org/10.15408/jti.v14i2.24024

Handayani, E. T., & Sulistiyawati, A. (2021). Analisis Sentimen Respon Masyarakat terhadap Kabar Harian Covid-19 pada Twitter Kementerian Kesehatan dengan Metode Klasifikasi Naive Bayes. *Jurnal Teknologi Dan Sistem Informasi (JTSI)*, *2*(3), 32–37. https://doi.org/10.33365/jtsi.v2i3.906

Handoko, W. T., Supriyanto, E., Purwadi, D. I., Budiarso, Z., & Listiyono, H. (2022). Klasifikasi Opini Pengguna Media Sosial Twitter Terhadap JNT Di Indonesia dengan Algoritma Decision Tree. *Jurnal Sains Komputer & Informatika (J-SAKTI)*, *6*(2), 790–799.

Herdhianto, A. (2020). *Sentiment Analysis Menggunakan Naïve Bayes Classifier (NBC) Pada Tweet Tentang Zakat*.

Karima, A. (2022). *PREDIKSI KINERJA MAHASISWA DALAM PERKULIAHAN DARING BERBASIS LEARNING MANAGEMENT SYSTEM (LMS) MENGGUNAKAN ALGORITMA NAÏVE BAYES*. UNIVERSITAS MUHAMMADIYAH KALIMANTAN TIMUR.

Kominfo. (2022). *Kominfo : Pengguna Internet di Indonesia 63 Juta Orang*. https://www.kominfo.go.id/index.php/content/detail/3415/Kominfo+:+Pengguna+Internet+di+Indonesia+63+Juta+Orang/0/berita_satker

Mahardika, Y. S., & Zuliarso, E. (2018). Analisis Sentimen Terhadap Pemerintahan Joko Widodo Pada Media Sosial Twitter Menggunakan Algoritma Naives Bayes. *Prosiding SINTAK 2018*, *2015*, 409–413.

Patro, R. (2021). *Cross-Validation: K Fold vs Monte Carlo*. Medium. https://towardsdatascience.com/cross-validation-k-fold-vs-monte-carlo-e54df2fc179b

Sari, D. N., Sari, D. N., Adelia, F., Rosdiana, F., Butar, B. B., & Hariyanto, M. (2020). Analisa Sentimen Terhadap Review Produk Kecantikan Menggunakan Metode Naive Bayes Classifier. *JIKA (Jurnal Informatika)*, *4*(3), 109. https://doi.org/10.31000/jika.v4i3.3086

Sari, F. V., & Wibowo, A. (2019). Analisis Sentimen Pelanggan Toko Online Jd.Id Menggunakan Metode Naïve Bayes Classifier Berbasis Konversi Ikon Emosi. *Jurnal SIMETRIS*, *10*(2), 681–686.

Susanti, N. A., & Walid, M. (2022). *Klasifikasi Data Tweet Ujaran Kebencian Di Media Sosial*. *6*(2), 538–543.

Suyanto, D. (2019). *Data Mining untuk Klasifikasi dan Klasterisasi Data* (Revisi). Informatika Bandung.

Wibowo, J. S., Semarang, U. S., Semarang, K., & Tengah, J. (2023). *PENERAPAN ALGORITMA NAÏVE BAYES CLASSIFICATION UNTUK KLASIFIKASI SENTIMENT TWEET TERHADAP PLATFORM STREAMING ILEGAL*. *2*, 225–233.

Yutika, C. H., Adiwijaya, A., & Faraby, S. Al. (2021). Analisis Sentimen Berbasis Aspek pada Review Female Daily Menggunakan TF-IDF dan Naïve Bayes. *Jurnal Media Informatika Budidarma*, *5*(2), 422. https://doi.org/10.30865/mib.v5i2.2845

Zhafira, D. F., Rahayudi, B., & Indriati, I. (2021). Analisis Sentimen Kebijakan Kampus Merdeka Menggunakan Naive Bayes dan Pembobotan TF-IDF

Berdasarkan Komentar pada Youtube. *Jurnal Sistem Informasi, Teknologi Informasi, Dan Edukasi Sistem Informasi*, *2*(1), 55–63. https://doi.org/10.25126/justsi.v2i1.24