

BAB IV

HASIL DAN PEMBAHASAN

4.1 Hasil Pengumpulan Data

Data yang telah melalui tahap pengambilan data akan diunggah ke Google Drive kemudian dihubungkan ke Google Collab menggunakan *library* yang terdapat pada Google Collab yaitu Pandas. Implementasi dari proses ini dapat dilihat pada potongan kode program di bawah ini.

```
import pandas as pd
df = pd.read_csv("/content/drive/MyDrive/skripsi/dataset.csv")
```

Gambar 4.1 Fungsi Pandas di Google Collab

Berikut merupakan tampilan visualisasi data yang telah di *import* ke dalam program.



	komentar	HS
0	disini kita melihat orang dewasa sprti anak an...	0
1	Ngapa ada boteng wkwkwk	0
2	sultan mau ngapain coba??	0
3	7:49 ngakak gini doang padahal	0
4	akhirnya 22 taun hidup dari konten ini nemu or...	0
...
9995	Sub nya cuma [Music] sepanjang video, lho kok ...	0
9996	Duh gangerti lagi kontennya gokil parah,	0
9997	Harus ya half naked gitu?	0
9998	RAY WIBU	0
9999	Harusnya mukanya yang muka indo min,maaf bukan...	0

10000 rows × 2 columns

Gambar 4.2 Tampilan visualisasi data mentah

4.2 Hasil Penelitian

4.2.1 Data Cleansing

Data yang telah di *import* kemudian akan dilakukan data *cleansing* untuk menghilangkan komponen yang tidak dibutuhkan.



	komentar	HS	cleaned
0	disini kita melihat orang dewasa sprti anak an...	0	disini kita melihat orang dewasa sprti anak an...
1	Ngapa ada boteng wkwkwk	0	Ngapa ada boteng wkwkwk
2	sultan mau ngapain coba??	0	sultan mau ngapain coba
3	7:49 ngakak gini doang padahal	0	ngakak gini doang padahal
4	akhirnya 22 taun hidup dari konten ini nemu or...	0	akhirnya taun hidup dari konten ini nemu org ...
...
9995	Sub nya cuma [Music] sepanjang video, lho kok ...	0	Sub nya cuma Music sepanjang video lho kok tum...
9996	Duh gangerti lagi kontennya gokil parah,	0	Duh gangerti lagi kontennya gokil parah
9997	Harus ya half naked gitu?	0	Harus ya half naked gitu
9998	RAY WIBU	0	RAY WIBU
9999	Harusnya mukanya yang muka indo min,maaf bukan...	0	Harusnya mukanya yang muka indo minmaaf bukan ...

Gambar 4. 3 Hasil proses data *cleansing*

Dapat dilihat pada gambar diatas untuk data yang telah melewati tahapan data *cleansing* kalimat yang awalnya memiliki komponen angka dan tanda baca telah dibersihkan.

4.2.2 Case Folding

Hasil dari proses *case folding* dapat dilihat pada gambar di bawah. Setelah melalui proses *case folding*, kalimat yang awalnya memiliki huruf kapital diubah menjadi huruf kecil (*lowercase*). Hasil dari proses *case folding* dapat dilihat pada gambar 4.4.

0	disini kita melihat orang dewasa sprti anak an...	disini kita melihat orang dewasa sprti anak an...
1	Ngapa ada boteng wkwkwk	ngapa ada boteng wkwkwk
2	sultan mau ngapain coba??	sultan mau ngapain coba
3	7:49 ngakak gini doang padahal	ngakak gini doang padahal
4	akhirnya 22 taun hidup dari konten ini nemu or...	akhirnya taun hidup dari konten ini nemu org ...
...
9995	Sub nya cuma [Music] sepanjang video, lho kok ...	sub nya cuma music sepanjang video lho kok tum...
9996	Duh gangerti lagi kontennya gokil parah,	duh gangerti lagi kontennya gokil parah
9997	Harus ya half naked gitu?	harus ya half naked gitu
9998	RAY WIBU	ray wibu
9999	Harusnya mukanya yang muka indo min,maaf bukan...	harusnya mukanya yang muka indo minmaaf bukan ...

10000 rows x 3 columns

Gambar 4. 4 Hasil proses *case folding*

4.2.3 *Tokenizing*

Hasil dari proses *tokenizing* dapat dilihat pada gambar di bawah. Setelah melalui proses *tokenizing*, data yang awalnya berbentuk kalimat dicacah menjadi kata-kata. Hasil dari proses *tokenizing* dapat dilihat pada gambar 4.5.

0	disini kita melihat orang dewasa sprti anak an...	[disini, kita, melihat, orang, dewasa, sprti, ...
1	Ngapa ada boteng wkwkwk	[ngapa, ada, boteng, wkwkwk]
2	sultan mau ngapain coba??	[sultan, mau, ngapain, coba]
3	7:49 ngakak gini doang padahal	[, ngakak, gini, doang, padahal]
4	akhirnya 22 taun hidup dari konten ini nemu or...	[akhirnya, , taun, hidup, dari, konten, ini, n...
...
9995	Sub nya cuma [Music] sepanjang video, lho kok ...	[sub, nya, cuma, music, sepanjang, video, lho,...
9996	Duh gangerti lagi kontennya gokil parah,	[duh, gangerti, lagi, kontennya, gokil, parah]
9997	Harus ya half naked gitu?	[harus, ya, half, naked, gitu]
9998	RAY WIBU	[ray, wibu]
9999	Harusnya mukanya yang muka indo min,maaf bukan...	[harusnya, mukanya, yang, muka, indo, minmaaf,...

10000 rows x 4 columns

Gambar 4. 5 Hasil proses *tokenizing*

4.2.4 Normalisasi

Hasil dari proses normalisasi dapat dilihat pada gambar di bawah ini. Setelah melalui proses normalisasi, kata yang awalnya merupakan kata gaul atau singkatan telah diproses menjadi kata aslinya. Hasil dari proses normalisasi dapat dilihat pada gambar 4.6.

tokenized	normalized
[disini, kita, melihat, orang, dewasa, sprti, ...	[disini, sekitar, melihat, orang-orang, dewasa...
[ngapa, ada, boteng, wkwkwk]	[mengapai, enggak ada, boteng, wkwkwk]
[sultan, mau, ngapain, coba]	[sultan, mau, mengapai, coba]
[, ngakak, gini, doang, padahal]	[mengakak, begini, doang, padahal]
[akhirnya, , taun, hidup, dari, konten, ini, n...	[akhirnya, tahun, hidup, dari, konten-konten, ...
...	...
[sub, nya, cuma, music, sepanjang, video, lho,...	[subhanallah, kata-katanya, percuma, music, se...
[duh, gangerti, lagi, kontennya, gokil, parah]	[duh, gangerti, lagi, kontennya, gila, parah]
[harus, ya, half, naked, gitu]	[seharusnya, kata-katanya, half, naked, begitu]
[ray, wibu]	[merayakan, wibu]
[harusnya, mukanya, yang, muka, indo, minmaaf,...	[seharusnya, mukanya, menyangka, mukanya, indo...

Gambar 4. 6 Hasil proses normalisasi

4.2.5 Stopwords Removing

Hasil dari proses *stopwords removing* dapat dilihat pada gambar di bawah ini. Setelah melewati proses *stopwords removing* data yang awalnya mengandung kata “dan”, “di”, “dari”, “yang”, dll telah dihapus. Hasil dari proses *stopwords removing* dapat dilihat pada gambar 4.7.

normalized	stops
[sekitar, orang-orang, dewasa, anaknya, anakny...	[sekitar, orang-orang, dewasa, anaknya, anakny...
[mengapai, enggak ada, boteng, wkwkwk]	[mengapai, enggak ada, boteng, wkwkwk]
[sultan, mengapai, coba]	[sultan, mengapai, coba]
[mengakak, doang]	[mengakak, doang]
[tahun, hidup, konten-konten, menemukan, orang...	[tahun, hidup, konten-konten, menemukan, orang...
...	...
[subhanallah, kata-katanya, music, video, loh,...	[subhanallah, kata-katanya, music, video, loh,...
[duh, gangerti, kontennya, gila, parah]	[duh, gangerti, kontennya, gila, parah]
[kata-katanya, half, naked]	[kata-katanya, half, naked]
[merayakan, wibu]	[merayakan, wibu]
[mukanya, menyangka, mukanya, indonesia, minma...	[mukanya, menyangka, mukanya, indonesia, minma...

Gambar 4. 7 Hasil proses *stopwords removing*

4.2.6 *Stemming*

Hasil dari proses *stemming* dapat dilihat pada gambar di bawah ini. Setelah melewati proses *stemming*, kata yang awalnya memiliki imbuhan seperti prefiks, sufiks, maupun konfiks yang terdapat pada dokumen akan dihilangkan.

stops	stems
[sekitar, orang-orang, dewasa, anaknya, anakny...	[sekitar, orang, dewasa, anak, anak, turut]
[mengapai, enggak ada, boteng, wkwkwk]	[kapai, enggak ada, boteng, wkwkwk]
[sultan, mengapai, coba]	[sultan, kapai, coba]
[mengakak, doang]	[akak, doang]
[tahun, hidup, konten-konten, menemukan, orang...	[tahun, hidup, konten, temu, orang, kayak, gue...
...	...
[subhanallah, kata-katanya, music, video, loh,...	[subhanallah, kata, music, video, loh, pokok, ...]
[duh, gangerti, kontennya, gila, parah]	[duh, gangerti, konten, gila, parah]
[kata-katanya, half, naked]	[kata, half, naked]
[merayakan, wibu]	[raya, wibu]
[mukanya, menyangka, mukanya, indonesia, minma...	[muka, sangka, muka, indonesia, minmaaf, maksu...

Gambar 4. 8 Hasil proses *stemming*

4.2.7 *Training Model*

Pada tahap *training model*, *library* yang digunakan berasal dari *scikit.learn*. Penggunaan *library* dari *scikit.learn* ini bertujuan untuk memudahkan proses *training model*. Data akan disimpan dalam 4 variabel yaitu *x_test*, *y_test*, *x_train* dan *y_train*. *x_train* untuk data bobot ulasan yang digunakan sebagai data latih. *y_train* untuk data label yang digunakan sebagai data latih. *x_test* untuk data ulasan

yang digunakan sebagai data uji. y_{test} untuk data label yang digunakan sebagai data uji. Pada training model ini menggunakan 80% data sebagai data latih dan 20% sebagai data uji. Hasil dari proses *training data* dan *confusion matrix* dapat dilihat pada dibawah.

```
f1 score hasil prediksi adalah:
0.7368421052631579
accuracy score hasil prediksi adalah:
0.995
preccision score hasil prediksi adalah:
1.0
recall score hasil prediksi adalah:
0.5833333333333334
```

Gambar 4. 9 Hasil proses *training data*

Tabel 4. 1 Tabel *Confusion Matrix*

	Actual Positive	Actual Negative
Predict Positive	14	0
Predict Negative	10	1976

Berdasarkan Tabel 4.1, ditunjukkan bahwa 1990 diantara 2000 data uji telah diklasifikasikan dengan benar, namun 10 data uji masih diprediksikan salah dengan kelas aktualnya adalah negatif dan hasil klasifikasi yang diberikan adalah kelas positif. Dari hasil confusion matrix dapat dihitung nilai *accuracy*, *precision*, *recall*, dan *f1-score* sebagai berikut:

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} = \frac{1976 + 14}{1976 + 14 + 10 + 0} = 0.995$$

$$Precision = \frac{TP}{TP + FP} = \frac{14}{14 + 0} = 1$$

$$Recall = \frac{TP}{TP + FN} = \frac{14}{14 + 10} = 0.583$$

$$F1 - score = \frac{2 * (Recall * Precision)}{(Recall + Precision)} = \frac{2 * (0.583 * 1)}{(0.583 + 1)} = 0.736$$

4.3 Pembahasan

Penulis menggunakan *library* dari *sklearn* yaitu *metrics* untuk melakukan evaluasi terhadap akurasi model klasifikasi yang dibuat. *Function* yang digunakan untuk melakukan evaluasi terhadap akurasi pemodelan yang telah dibuat adalah *f1_score*, *precision_score*, *confusion_matrix*, dan *accuracy_score*. Dengan menggunakan *function* ini akan memudahkan penulis untuk melihat hasil dari masing-masing parameter yang telah ditentukan.

Hasil dari klasifikasi ujaran kebencian pada komentar video YouTube Cretivox menggunakan algoritma SVM dengan pembobotan data menggunakan metode TF-IDF dengan jumlah data latih 80% dan 20% data uji menghasilkan hasil akurasi sebesar 99%. *Precision score* yang dihasilkan sebesar 100%, *f1 score* sebesar 73,68%, sedangkan *recall score* sebesar 58,3% . Dengan nilai akurasi yang tinggi yaitu 99%, dapat disimpulkan bahwa algoritma *Support Vector Machine* dengan pembobotan data menggunakan metode TF-IDF memiliki akurasi yang tinggi untuk klasifikasi ujaran kebencian.