

**Data Mining**

**GEMASTIK 13**

**Pembangunan Model Analisis Sentimen Komentar Masyarakat di Youtube  
Terhadap Kebijakan *New Normal***

**Disusun Oleh:**

Regy Saputra

Yuliansyah Ibrahim

TIM DataMiningUPI1

FAKULTAS PENDIDIKAN MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS PENDIDIKAN INDONESIA

2020

## Daftar Isi

<b>1.</b>	<b>Latar Belakang</b>	<b>1</b>
<b>2.</b>	<b>Tujuan dan Manfaat</b>	<b>2</b>
<b>3.</b>	<b>Batasan</b>	<b>2</b>
<b>4.</b>	<b>Metode</b>	<b>2</b>
4.1.	Dataset.	3
4.2.	TF-IDF.	3
4.3.	Logistic Regression.	4
<b>5.</b>	<b>Desain dan Implementasi</b>	<b>4</b>
5.1.	Desain.	4
5.2.	Pengambilan dan Pelabelan Data.	5
5.3.	Praproses	7
5.4.	Pelatihan dan Pengujian.	7
<b>6.</b>	<b>Analisis</b>	<b>8</b>
<b>7.</b>	<b>Kesimpulan</b>	<b>9</b>
	<b>Daftar Pustaka</b>	<b>10</b>
	<b>Dokumentasi</b>	<b>11</b>

## 1. Latar Belakang

Sejak bulan Maret 2020, organisasi kesehatan dunia WHO telah mengumumkan kejadian covid-19 sebagai kejadian pandemi dunia. Sejak saat itu, pemerintah Indonesia telah menerapkan kebijakan PSBB (Pembatasan Sosial Berskala Besar) untuk menekan laju penyebaran virus covid-19 di dalam negeri. Namun kebijakan PSBB yang terlalu lama telah berdampak terhadap perkembangan ekonomi nasional yang memburuk. Sebagai respon keadaan tersebut, mulai dari tanggal 2 Juni 2020, pemerintah secara bertahap mengimplementasikan kebiasaan hidup yang baru atau disebut kebijakan status “New Normal”, sebagai upaya agar masyarakat dapat kembali produktif namun tetap aman dari Covid-19. Menggunakan masker saat bepergian, menjaga jarak saat beraktivitas adalah sesuatu yang sebelumnya kita tidak bayangkan akan menjadi suatu kebiasaan yang harus dilakukan saat ini. Banyak pro dan kontra yang terjadi di masyarakat mengenai kebijakan tersebut. Dalam kondisi pandemi yang berbahaya seperti pada saat ini, kebijakan pemerintah sangat menentukan bagaimana kondisi kesehatan maupun ekonomi nasional. Pemerintah harus mampu melakukan evaluasi dari kebijakan yang diambil secara cepat sehingga dapat memastikan kembali kebijakan yang sesuai dengan kebutuhan. Proses evaluasi kebijakan dapat dilakukan dengan menganalisis bagaimana respon dari masyarakat terhadap kebijakan pemerintah.

Informasi media sosial dapat digunakan sebagai salah satu sumber data yang dapat digunakan untuk mengevaluasi kebijakan pemerintah. Berdasarkan laporan penelitian terbaru oleh lembaga *We Are Social*, pada tahun 2020 disebutkan bahwa jumlah pengguna internet di Indonesia mencapai 175,4 juta atau 64% dari total penduduk Indonesia adalah pengguna internet aktif [1]. Dalam laporan penelitian tersebut juga diketahui bahwa 160 juta penduduk Indonesia adalah pengguna aktif media sosial dan media sosial yang paling banyak digunakan adalah youtube. Dari laporan tersebut menunjukkan bahwa masyarakat Indonesia mendapat informasi paling banyak adalah dari youtube. Selain itu, platform youtube juga menyediakan fitur untuk menulis komentar terhadap setiap tayangan atau konten yang disajikan dalam videonya. Hal ini memudahkan masyarakat untuk menyampaikan atau menuliskan tanggapan terhadap setiap informasi yang ditayangkan. Hal ini merupakan hal yang positif bagi pemerintah karena pemerintah dapat memanfaatkan informasi tersebut untuk

mengevaluasi setiap kebijakan yang diambil dengan melakukan analisis terhadap setiap komentar yang berhubungan dengan kebijakan pemerintah, khususnya yang berkaitan dengan kebijakan yang berhubungan dengan penanganan kejadian covid-19. Untuk saat ini kebijakan yang sedang dilakukan dan perlu segera dievaluasi adalah kebijakan mengenai status new normal dalam kondisi covid-19.

Salah satu disiplin ilmu dalam bidang data mining yang dapat digunakan untuk melakukan proses analisis data text adalah Natural Language Processing (NLP). Metode-metode dalam NLP dapat digunakan untuk menganalisa pengetahuan di dalam teks sehingga teks tersebut menjadi sebuah informasi yang berguna, salah satunya adalah untuk menganalisis sentimen dari setiap informasi yang berupa text. Dalam penelitian ini akan dilakukan pembangunan model untuk mengklasifikasi sentimen komentar yang ada di video berita yang memberitakan kebijakan pemerintah khususnya yang membahas mengenai kebijakan status new normal. Kelas sentimen yang akan digunakan dalam proses analisis terbagi menjadi 3 kategori yaitu positif, negatif dan netral. Selanjutnya berdasarkan hasil analisis tersebut akan memberikan informasi kepada pemerintah sebagai bahan untuk evaluasi kebijakan yang berhubungan dengan covid-19.

## **2. Tujuan dan Manfaat**

Penelitian yang dilakukan bertujuan untuk menganalisis sentimen masyarakat terhadap kebijakan berkaitan dengan new normal dari pemerintah. Analisis yang dilakukan berdasarkan pada komentar youtube. Komentar yang dianalisis dibagi menjadi 3 kelompok sentimen, yaitu sentimen positif, netral, dan negatif. Dengan adanya analisis sentimen ini diharapkan dapat membantu pemerintah mengetahui umpan balik serta sebagai tolak ukur tingkat kepuasan masyarakat terhadap kebijakan yang diterapkan.

## **3. Batasan**

Pada penelitian ini data yang digunakan untuk melakukan analisis sentimen masyarakat hanya terbatas pada data komentar youtube di lima video berita.

## **4. Metode**

Penelitian ini mengklasifikasikan data menjadi tiga kelas, yaitu kelas positif,

kelas negatif dan kelas netral. Data yang dipakai adalah data komentar pengguna Youtube pada lima video berita mengenai kebijakan *New Normal*.

ID Video	Judul Video	Channel
V9K9gONlh5g	Sorotan: Indonesia Bersiap Menuju New Normal?	KOMPASTV
6D74v9ABSIg	[FULL] Jokowi: New Normal, Kita Harus Berkompromi Dengan Covid	KOMPASTV
VzXKijeWN4U	Terdampak Covid-19, RI Siap Masuki Era New Normal?	CNBC Indonesia
dFPepr6T7y0	Sorotan: Indonesia Bersiap Jalani New Normal	KOMPASTV
ozm7lCEK328	Catat! Mulai 1 Juni Kebijakan New Normal, Jurus Pamungkas Selamatkan Ekonomi?   tvOne	tvOneNews

Tabel 1: Video berita *New Normal*

#### 4.1. Dataset

Untuk mendapatkan data komentar di Youtube, kami memanfaatkan Youtube Data API yang dikembangkan oleh Youtube sendiri. Kami membuat *script* sederhana di Python untuk mengambil dan mengumpulkan seluruh komentar yang ada pada lima video tersebut, kemudian data komentar dibersihkan dan diberikan label sentimen secara manual. Di dalam dataset terdapat 597 komentar negatif, 543 komentar netral, dan 320 komentar positif terhadap kebijakan *New Normal*. Dari 1460 komentar, sebanyak 1280 data digunakan untuk training dan 120 digunakan untuk testing.

#### 4.2. TF-IDF

TF-IDF merupakan sebuah metode *feature extraction* yang digunakan untuk menghitung nilai yang mencerminkan relevannya sebuah kata dalam sebuah dokumen di dalam korpus [2]. Pada dasarnya, TF-IDF bekerja dengan menentukan frekuensi relatif kata dalam dokumen tertentu (TF) dibandingkan dengan proporsi kebalikan dari kata tersebut di seluruh korpus dokumen (IDF). Dengan IDF, jika sebuah token langka muncul di dua dokumen, maka token tersebut memiliki arti penting bagi tiap dokumen[3]. IDF memberi beban pada token  $t$  pada kumpulan dokumen  $U$ , yang dapat dihitung seperti berikut:

$$IDF(t) = \frac{N}{n(t)}$$

dimana  $\frac{n(t)}{N}$  adalah frekuensi dari  $t$  di dalam  $U$  dan  $\frac{N}{n(t)}$  adalah inverse frekuensi. Sehingga total beban TF-IDF untuk sebuah token di dalam dokumen adalah:

$$TF-IDF = TF * IDF$$

Secara intuitif, kalkulasi ini menentukan seberapa relevan suatu kata dalam dokumen tertentu. Kata-kata yang umum dalam satu atau sekelompok kecil dokumen cenderung memiliki angka TF-IDF yang lebih tinggi daripada kata-kata umum seperti artikel dan preposisi.

#### 4.3. Logistic Regression

Metode *logistic regression* adalah suatu model analisis statistika yang mendeskripsikan hubungan antara variabel respons yang memiliki dua kategori atau lebih dengan satu atau lebih variabel bebas berskala kategori atau interval [4]. *logistic regression* dengan tiga prediktor variabel dapat digambarkan secara matematika dengan [5]:

$$\begin{aligned} \text{logit}(P) = \ln \left( \frac{P}{1-P} \right) = & \beta_0 + C_1X + C_2Y + C_3Z \\ & + C_4X^2 + C_5Y^2 + C_6Z^2 \\ & + C_7XY + C_8XZ + C_9YZ \end{aligned}$$

dimana  $\beta$  adalah intercept,  $X$ ,  $Y$ , dan  $Z$  adalah variabel predictor, dan  $C_i$  adalah koefisien, dan  $P$  adalah probabilitas akan terjadinya peristiwa tersebut berdasarkan perhitungan model. Hal ini membuat *logistic regression* memberi hasil yang bagus pada dataset dengan TF-IDF [3].

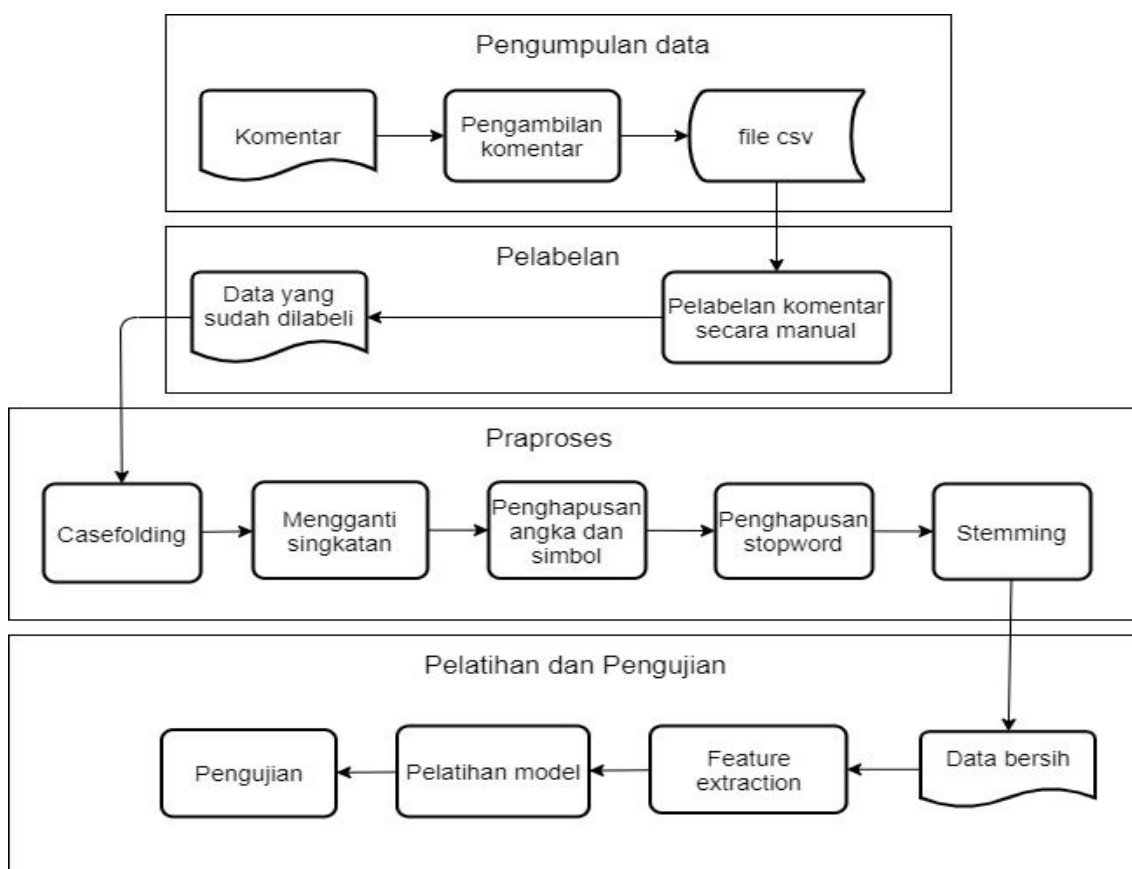
### 5. Desain dan Implementasi

#### 5.1. Desain

Alur penelitian yang kami lakukan dapat dibagi menjadi 4 bagian utama, yaitu pengumpulan data, pelabelan, praproses, serta pelatihan dan pengujian. Pada tahapan pengumpulan data, dilakukan proses pengambilan data komentar menggunakan API. Selanjutnya pada tahap pra proses, kami melakukan berbagai

proses, mulai dari *casefolding*, mengganti singkatan, penghapusan simbol dan angka, penghapusan stopwords, dan yang terakhir adalah proses stemming. Semua proses tersebut dilakukan untuk menghasilkan data yang bersih.

Tahap terakhir adalah proses pelatihan dan pengujian model dengan menggunakan data hasil dari tahap sebelumnya. Proses feature extraction menggunakan TF-IDF dilakukan pada data untuk mengubah teks menjadi matriks yang berisi informasi tentang relevansi kata, sehingga dapat digunakan untuk melatih model. Setelah dilatih, model diuji dengan dataset uji. Alur dari penelitian dapat dilihat pada gambar 2 dibawah.



Gambar 2: Diagram alur penelitian

## 5.2. Pengambilan dan Pelabelan Data

Data yang digunakan pada penelitian ini adalah data *thread* komentar dalam 5 video yang memberitakan tentang new normal. *Thread* komentar adalah komentar yang di dalamnya dapat memuat balasan komentar lainnya. Data *thread* komentar tersebut diambil menggunakan API yang dikeluarkan oleh Youtube sendiri. Dari 5

video tersebut, terdapat sebanyak 8233 *thread* komentar yang dapat kami diambil. Data yang telah diambil lalu dibersihkan dengan membuang komentar yang berisi link dan juga komentar yang merupakan kalimat pendek yang terdiri kurang dari 5 kata. Proses ini menyisakan sebanyak 4913 komentar pada data.

Judul Video	Jumlah <i>Thread</i> Komentar
Sorotan: Indonesia Bersiap Menuju New Normal?	235
[FULL] Jokowi: New Normal, Kita Harus Berkompromi Dengan Covid	1.681
Terdampak Covid-19, RI Siap Masuki Era New Normal?	1.101
Sorotan: Indonesia Bersiap Jalani New Normal	350
Catat! Mulai 1 Juni Kebijakan New Normal, Jurus Pamungkas Selamatkan Ekonomi?   tvOne	4.866

Tabel 2: Jumlah *thread* Komentar pada tiap video

Data yang sudah dibersihkan lalu diacak dan diambil sebanyak 1460 komentar untuk kemudian diberi label sentimen secara manual. Sentimen komentar dibagi menjadi 3 kelompok, yaitu komentar yang tidak suka dengan adanya kebijakan *new normal*, komentar yang netral dan tidak peduli, dan yang terakhir adalah komentar yang menerima dan mendukung adanya *new normal*. Seperti yang telah disebutkan sebelumnya, dari 1460 data yang telah diberi label, terdapat sebanyak 597 komentar bersentimen negatif, 543 komentar bersentimen netral, dan 320 komentar bersentimen positif.

Sentimen	Jumlah Komentar	Sampel Komentar
Negatif	597	Tinggal bilang pasrah aja kok repot...udah ga sanggup kasih makan rakyat...
Netral	543	Tetap waspada dan pakai masker jika pergi kemanapun
Positif	320	Semoga lancar dan tidak menyulitkan




		rakyat mencari nafkah
--	--	-----------------------

Tabel 3: Distribusi sentimen di dataset

### 5.3. Praproses

Salah satu karakteristik pada pesan dan teks di platform media sosial adalah penggunaan kata yang tidak baku. Selain itu, kata yang digunakan juga disingkat seperti kata 'tidak' bisa ditulis menjadi 'tdk', 'g', 'gk', 'ga', 'gak', 'ngga', 'nggak', 'engga', dan juga 'enggak'. Hal ini sangat berpengaruh pada kinerja model karena kalimat 'tdk baik', 'ga baik', dan 'tidak baik' dapat dihitung sebagai kalimat yang berbeda oleh mesin meskipun pada dasarnya kalimat tersebut memiliki makna yang sama. Maka pada tahapan preprocessing kami menyertakan proses untuk mengubah beberapa kata yang umum disingkat menjadi bentuk bakunya.

Untuk proses penghilangan *stopword*, dan *stemming*, kami menggunakan *library python Sastrawi* yang di dalamnya terdapat daftar kata *stopword* dan fungsi untuk melakukan *stemming* pada Bahasa Indonesia. Berikut adalah perbandingan teks komentar sebelum dan sesudah *preprocessing*.

Teks Asli	Sesudah Praproses
Tatanan baru.....Seperti kalimatnya Sunda Empire... Orangnya ditangkap tp kalimatnya dipakai...wkwkwkwwk	tatanan baru kalimat sunda empire orang ditangkap kalimat pakai
Sekarang saya malu , kenapa dulu terlalu membanggakan beliau .. 	sekarang malu kenapa dulu terlalu bangga beliau
Mudah"Han dengan adanya new normal masyarakat yg perekonomiannya melemah segera bangkit kembali..amin..amin..amin	mudah han ada masyarakat ekonomi lemah segera bangkit amin amin amin

Tabel 4: Perbandingan teks komentar sebelum dan sesudah praproses

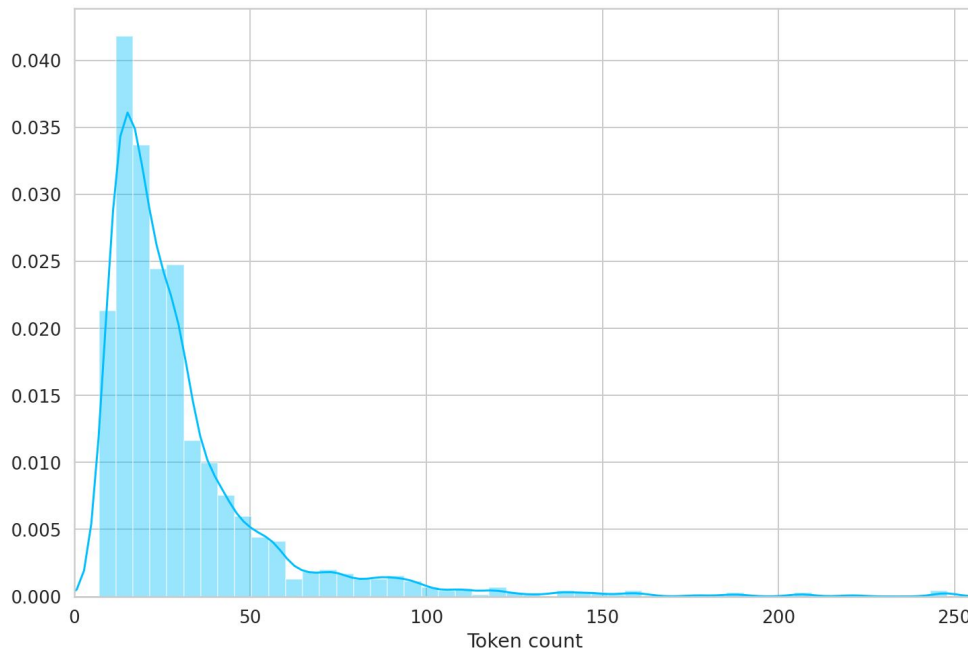
### 5.4. Pelatihan dan Pengujian

Setelah dilakukan praproses, data yang dihasilkan diubah ke bentuk matriks TF-IDF dengan menggunakan fungsi *TfidfVectorizer* dari *library sklearn*. Nilai df minimum adalah 2 dan nilai df maksimum adalah 0,2. Untuk membangun model

*logistic regression* di Python, kami menggunakan implementasi *logistic regression* dari *library sklearn* dengan iterasi maksimal 600 dan parameter class weight diubah menjadi *balanced*. Kemudian model dilatih dengan dataset *training* yang berisi 1280 komentar. Setelah itu model diuji dengan dataset *test* berisi 120 komentar yang terdiri dari masing-masing 60 label sentimen negatif, netral, dan positif. Algoritma klasifikasi lainnya, yaitu *Multinomial naïve Bayes* dan *SVM* digunakan sebagai perbandingan nilai akurasi model.

## 6. Analisis

Tujuan dari penelitian ini adalah membangun model untuk mengenali kelompok opini berdasarkan komentar di Youtube. Kemudian model digunakan untuk melakukan klasifikasi apakah sebuah komentar termasuk dalam class C, dimana C adalah negatif, netral, atau positif. Dataset terdiri 1460 komentar pengguna youtube. Rata-rata jumlah kata di setiap komentar adalah sebanyak 150 kata. Distribusi jumlah kata dapat dilihat pada gambar 2 dibawah.



Gambar 2: Plot distribusi jumlah kata

Pengujian dilakukan dengan mengukur akurasi dan f1-score dari hasil perhitungan *logistic regression*, *multinomial Naïve Bayes*, *support vector machine*. Adapun total data komentar yang digunakan untuk pengujian adalah sebanyak 120 komentar yang terdiri dari 60 komentar negatif, 60 komentar netral, dan 60 komentar positif. Hasil

pengujian ditunjukkan oleh gambar dibawah. Pengujian klasifikasi diukur dengan nilai akurasi dan f1-score yang diperoleh dengan membandingkan tiap komentar yang telah diberi label dengan hasil perhitungan dari model.

Model/Classifier	Feature	Akurasi	F1-score		
			Negatif	Netral	Positif
<i>Logistic Regression</i>	Bigram TF-IDF	64,4	64,7	60,7	67,9
	Unigram TF-IDF	62,7	62,8	58,7	66,7
<i>SVM</i>	Bigram TF-IDF	62,7	66,7	59,1	60,9
	Unigram TF-IDF	59,4	61,7	55,2	61,1
<i>Multinomial Naïve Bayes</i>	Bigram TF-IDF	63,3	68	58,4	61,7
	Unigram TF-IDF	61,6	64,1	57,7	62,4

Tabel 5: Perolehan akurasi dan f1-score

Pada tabel 5 diatas, terlihat bahwa nilai akurasi terbaik jatuh pada model *classifier logistic regression* dengan fitur *bigram* TF-IDF yang mendapatkan nilai akurasi sebesar 64,4% . Skor F1 pada label netral lebih kecil dari label lainnya. Hal ini mungkin disebabkan oleh komentar yang bersentimen netral tidak memiliki pola tertentu dan sulit ditebak.

## 7. Kesimpulan

Berdasarkan analisis di atas, dapat disimpulkan model klasifikasi sentimen yang paling cocok pada kasus ini adalah dengan menggunakan *logistic regression*. *Logistic regression* mampu mengatasi masukan data sparse, sehingga algoritma ini sangat cocok untuk data teks yang menggunakan TF-IDF sebagai feature extraction-nya. Meskipun begitu, dataset yang digunakan tergolong kecil, yaitu sebanyak 1460 komentar. Pengujian yang dilakukan menggunakan dataset sekecil ini tidak dapat dijadikan tolak ukur untuk topik besar seperti analisis sentimen. Pengumpulan data merupakan hambatan terbesar yang ada pada topik analisis sentimen media sosial .

## DAFTAR PUSTAKA

- [1] Tri Haryanto Agus. (2020, 20 Februari). *Riset: Ada 175,2 Juta Pengguna Internet di Indonesia*. Dikutip 12 September 2020. <https://inet.detik.com/>.
- [2] Rajaraman, A., & Ullman, J. (2011). Data Mining. In Mining of Massive Datasets (pp. 1-17). Cambridge: Cambridge University Press.
- [3] Rustam, F.; Ashraf, I.; Mehmood, A.; Ullah, S.; Choi, G.S. (2019). Tweets Classification on the Base of Sentiments for US Airline Companies. Entropy.
- [4] Hosmes, D.W. & S. Lemeshow. (1989). Applied Logistic regression. John Wiley and Sons, New York.
- [5] Zhao, L.; Chen, Y.; Schaffner, D.W. (2001). Comparison of logistic regression and linear regression in modeling percentage data. Appl. Environ. Microbiol.

# DOKUMENTASI

	A	B	C	D	E
966	UgwHAU6L4mH Timbul Buntuk1	SEBELUM VIRUS KORONA TERJADI EKONOMI .SUDAH AMBRUK....VIRUS KORONA JADI KAMBING HITAM ..DIBUAT ALASAN BAHWA VIRUS KORONA LAH .PENYEBAB EKONOMI ... SERIBU ALASAN .AGAR TERHINDAR DARI KESALAHAN...EKONOMI AMBRUK ..SEBELUM VIRUS KORONA .TERJADI .....JURUS PAMUNGKAS . HANYA DIBUAT AGAR MENGHINDAR DARI .KETIDAK MAMPUAN DALAM MENINGKATKAN EKONOMI	-1		Negatif
967	UgyCCsp1VTgT Hjjrah Tv	Yg di disimpulkan itu warga asing asing pa .bukan masyarakat ... Jgn dikambing hitamkan terus masyarakat ... yg berbahaya itu saat WNA bebas masuk ke Indonesia ... catat itu	-1		Negatif
968	Ugyz5Zm--P1oe kinan ario	Kok nggak ada yg fokus ke reporter nya.. cakep loh..	0		
969	Ugy7jNAa9uuFfi Lutfi Lutti	Kita saling percaya aja. Pasti tindakan ini udah dipikirkan matang" dari pemerintah..	1		Positif
970	Ugw0hTEzoq-5Z Sabrina Ayu	Inovasi bangsa Indonesia jgn New Normal harusnya ORHIBA ( orde hidup baru ) protokolnya pke cara peradaban nabi muhammad.. click ORHIBA	0		
971	UgyH5YFwgAlo Aziz deep	SUNDA EMPIRE naon eta???? akibat gak nyetor/reset ulang	0		
972	lInwen7Wd6aW Kevin Efendi	Saya setuju pak Jokowi. Tapi tolong berita2 di tv jangan menyiarkan atau memberitakan masalah corona dan jangan memberitakan masalah berita hnak biar rakyat adem ayem tentram	1		Positif

Dokumentasi 1: Memberi label sentimen pada komentar

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report

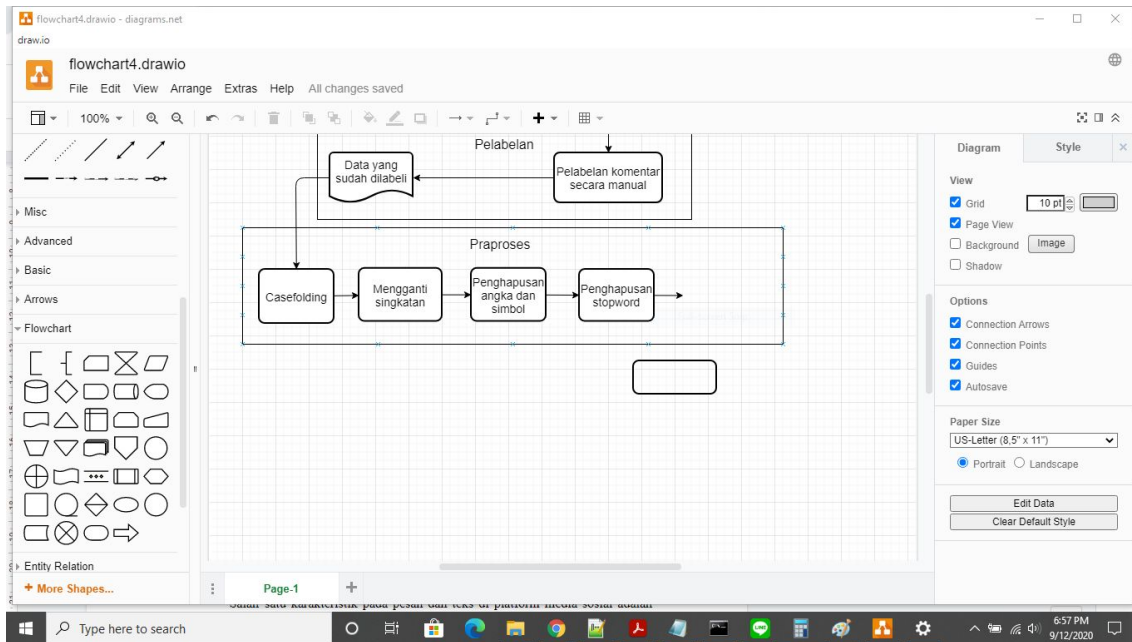
# Training
rgs.fit(X_train, y_train.values.ravel())

# Testing
y_pred = rgs.predict(X_test)
rgs.score(X_test, y_test)

print(metrics.classification_report(y_test, y_pred, digits=3))
```

	precision	recall	f1-score	support
-1.0	0.579	0.733	0.647	60
0.0	0.654	0.567	0.607	60
1.0	0.731	0.633	0.679	60
accuracy			0.644	180
macro avg	0.655	0.644	0.644	180
weighted avg	0.655	0.644	0.644	180

Dokumentasi 2: Membuat kode di Jupyter Notebook



### Dokumentasi 3: Mancang alur penelitian

Analisis Sentimen - Gemastik

komenter untuk kemudian diberi label sentimen secara manual. Sentimen komenter dibagi menjadi 3 kelompok, yaitu komenter yang tidak suka dengan adanya kebijakan *new normal*, komenter yang netral dan tidak peduli, dan yang terakhir adalah komenter yang menerima dan mendukung adanya *new normal*. Seperti yang telah disebutkan sebelumnya, dari 1460 data yang telah diberi label, terdapat sebanyak 597 komenter bersentimen negatif, 543 komenter bersentimen netral, dan 320 komenter bersentimen positif.

Sentimen	Jumlah Komenter	Sampel Komenter
Negatif	597	Tinggal bilang pasrah aja kok repot...udah ga sanggup kasih makan rakyat...
Netral	543	Tetap waspada dan pakai masker jika pergi kemanapun
Positif	320	Semoga lancar dan tidak menyulitkan rakyat mencari nafkah

**5.3. Preprocessing**

Pada

Salah satu karakteristik pada pesan dan teks di platform media sosial adalah penggunaan kata yang tidak baku. Selain itu, kata yang digunakan juga disingkat

### Dokumentasi 4: Menyusun makalah