

**ANALISIS SENTIMEN PADA PENGGUNAAN *HASTAG*
COVID – 19 DI MEDIA SOSIAL *TWITTER***

SKRIPSI



sebagai salah satu syarat untuk memperoleh gelar Sarjana Sains Terapan (S.ST)
di Program Studi Teknik Informatika
Jurusan Teknologi Informasi

Oleh

Yuniar Fabi Putra

E41171845

**PROGRAM STUDI TEKNIK INFORMATIKA
JURUSAN TEKNOLOGI INFORMASI
POLITEKNIK NEGERI JEMBER**

202

**ANALISIS SENTIMEN PADA PENGGUNAAN *HASTAG*
COVID – 19 DI MEDIA SOSIAL *TWITTER***

SKRIPSI



sebagai salah satu syarat untuk memperoleh gelar Sarjana Sains Terapan (S.ST)
di Program Studi Teknik Informatika
Jurusan Teknologi Informasi

Oleh

Yuniar Fabi Putra

E41171845

**PROGRAM STUDI TEKNIK INFORMATIKA
JURUSAN TEKNOLOGI INFORMASI
POLITEKNIK NEGERI JEMBER
2021**

**KEMENTERIAN PENDIDIKAN DAN KEBUDAYAAN
POLITEKNIK NEGERI JEMBER
JURUSAN TEKNOLOGI INFORMASI**

Analisis Sentimen Penggunaan Hastag COVID-19 di Media Sosial Twitter

Yuniar Fabi Putra (E41171845)


Telah Diuji pada Tanggal 9 Juli 2021
dan Dinyatakan Memenuhi Syarat

Ketua Penguji,



I Putu Dody Lesmana, ST, MT
NIP. 19790921 200501 1 001

Sekretaris Penguji,



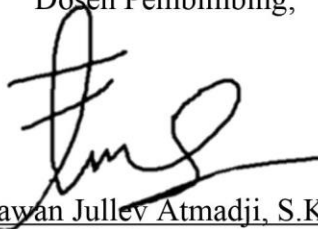
Ery Setiyawan Jullev Atmadji, S.Kom, M.Cs
NIP. 19890710 201903 1 010

Anggota Penguji,



Hermawan Arief Putranto, ST, MT
NIP. 19830109 201803 1 001

Dosen Pembimbing,



Ery Setiyawan Jullev Atmadji, S.Kom, M.Cs
NIP. 19890710 201903 1 010

Mengesahkan

Kepada Jurusan Teknologi Informasi



Hendra Yuhit Riskiawan, S.Kom, M.Cs
NIP. 19830203 200604 1 003

SURAT PERNYATAAN

Saya yang bertanda tangan di bawah ini :

Nama : Yuniar Fabi Putra

NIM : E41171845

Menyatakan dengan dengan sebenar-benarnya bahwa segala pernyataan dalam Laporan Skripsi saya yang berjudul “Analisis Sentimen Pada Penggunaan *Hashtag* COVID-19 di Media Sosial *Twitter*” merupakan gagasan dan hasil karya saya sendiri dengan arahan komisi pembimbing, dan belum pernah diajukan dalam bentuk apa pun pada perguruan tinggi mana pun.

Semua data dan informasi yang digunakan telah dinyatakan secara jelas dan dapat diperiksa kebenarannya. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan dari penulis dari penulis lain telah disebutkan dalam naskah dan dicantumkan dalam daftar pustaka di bagian akhir Laporan Skripsi ini.

Jember, 9 Juli 2021



Yuniar Fabi Putra

E41171845



**PERNYATAAN
PERSETUJUAN PUBLIKASI
KARYA ILMIAH UNTUK KEPENTINGAN
AKADEMIS**

Yang bertanda tangan dibawah ini, saya:

Nama : Yuniar Fabi Putra
NIM : E41171845
Program Studi : Teknik Informatika Kampus
Bondowoso
Jurusan : Teknologi Informasi

Demi mengembangkan Ilmu Pengetahuan, saya menyetujui untuk memberikan kepada UPT. Perpustakaan Politeknik Negeri Jember, Hak Bebas Royalti Non-Eksklusif (Non-Exclusive Royalty Free Right) atas Karya Ilmiah berupa **Laporan Skripsi** saya yang berjudul:

**ANALISIS SENTIMEN PADA PENGGUNAAN *HASTAG* COVID-19 DI
MEDIA SOSIAL *TWITTER***

Dengan Hak Bebas Royalti Non-Eksklusif ini UPT. Perpustakaan Politeknik Negeri Jember berhak menyimpan, mengalih media atau format, mengelola dalam bentuk Pangkalan Data (Database), mendistribusikan karya dan menampilkan atau mempublikasikannya di Internet atau media lain untuk kepentingan akademis tanpa perlu meminta ijin dari saya selama tetap mencantumkan nama saya sebagai penulis atau pencipta.

Saya bersedia untuk menanggung secara pribadi tanpa melibatkan pihak Politeknik Negeri Jember, Segala bentuk tuntutan hukum yang timbul atas Pelanggaran Hak Cipta dalam Karya ilmiah ini.

Demikian pernyataan ini saya buat dengan sebenarnya

Dibuat di : Jember
Pada Tanggal : 9 Juli 2021
Yang menyatakan,



Nama : Yuniar Fabi Putra
NIM : E41171845

MOTTO

“Urep Iku Masio Lunyu Kudu dipenek”

(Emha Ainun Nadjib)

PERSEMBAHAN

Dengan mengucapkan Bismillahirrahmanirrahim, sebagai rasa syukur atas segala limpahan rahmat, rizqi dan karunia yang diberikan oleh Allah SWT dalam pengerjaan Laporan Skripsi ini sehingga dapat terselesaikan dengan sebaik-baiknya dan tepat waktu. Laporan Skripsi ini dengan rasa bangga dipersembahkan sebagai bentuk rasa terima kasih kepada:

1. Kedua orang tua yang amat sangat saya cintai, Ibu Yatini yang telah melahirkan saya ke dunia dan menjadi seorang ibu yang sangat berbahagia dalam mendidik dan mengasuh putramu ini dengan segala doa-doa mu, serta Bapak Fatholah yang selalu menjadi seorang kepala keluarga yang sangat memperhatikan pendidikan dan kebahagiaan putramu ini.
2. Kedua kakak kandungku yang selalu mendukung adikmu ini untuk menjadi teladan terbaik dan menjadi kebanggaan bagi keluarga.
3. Bapak Ery Setiyawan Jullev Atmadji, S.Kom, M.Cs selaku Dosen Pembimbing yang telah memberikan arahan dan pembimbingan terbaik dalam mendampingi saya selama proses pengerjaan Skripsi.
4. Seluruh guru-guru yang pernah mengajarku di TK, SD, SMP dan SMA, serta dosen-dosen yang mengajarku selama kuliah di Politeknik Negeri Jember, terima kasih atas jasa-jasamu dalam memberikan edukasi dan motivasi yang luar biasa
5. Teman-teman Teknik Informatika 2017 Kampus Bondowoso yang selalu memberikan support dan kisah terbaik selama 4 tahun masa kuliah dalam menggapai gelar Sarjana ini.
6. Seluruh pihak dibelakangku yang tidak aku sebutkan satu persatu, terima kasih semoga kalian selalu diberkati dan diberi rahmat oleh Allah SWT, amin.

**Analisis Sentimen Pada Penggunaan *Hashtag* COVID-19 di Media Sosial
Twitter** (*Sentiment Analysis On The Use Of The Hashtag Covid-19 On Social
Media Twitter*)

Yuniar Fabi Putra
Study Program of Informatics Engineering
Majoring of Information Technology
Program Studi Teknik Informatika
Jurusan Teknologi Informasi

ABSTRACT

During this COVID-19 pandemic, in Indonesia, many users use Twitter to provide statements or opinions about COVID-19 by using the COVID-19 hashtag and making a tweet that tends to contain positive or negative opinions. The trend of tweets with the COVID-19 hashtag, can be known by opinion analysis or sentiment analysis. Sentiment analysis is an attempt to see the opinion or tendency of negative or positive sentiment based on the text of the tweet that can be matched with the topic being searched for. Therefore we need a classification that can analyze sentiment, especially tweets in Indonesian. The research was conducted using a classification method using Logistic Regression with TF-IDF word weighting calculations. By using the Logistic Regression classification, the score accuracy is 74% and the results of the system evaluation are 74.5% precision, 73% recall, and 73% F-Measure.

Keywords: *Sentiment Analysis, Logistic Regression, Covid-19*

Analisis Sentimen Pada Penggunaan *Hashtag* COVID-19 di Media Sosial *Twitter*

Yuniar Fabi Putra

Program Studi Teknik Informatika
Jurusan Teknologi Informasi

ABSTRAK

Di masa pandemi COVID-19 ini, di Indonesia banyak pengguna yang menggunakan Twitter untuk memberikan pernyataan atau opini tentang COVID-19 dengan menggunakan hashtag COVID-19 dan membuat tweet yang cenderung berisi opini positif atau negatif. Tren tweet dengan hashtag COVID-19, dapat diketahui dengan analisis opini atau analisis sentimen. Analisis sentimen adalah upaya untuk melihat pendapat atau kecenderungan sentimen negatif atau positif berdasarkan teks tweet yang dapat dicocokkan dengan topik yang dicari. Oleh karena itu diperlukan suatu klasifikasi yang dapat menganalisis sentimen khususnya tweet dalam bahasa Indonesia. Penelitian ini dilakukan dengan menggunakan metode klasifikasi menggunakan Regresi Logistik dengan perhitungan pembobotan kata TF-IDF. Dengan menggunakan klasifikasi Regresi Logistik, akurasi skor adalah 74% dan hasil evaluasi sistem adalah presisi 74,5%, recall 73%, dan F-Measure 73%.

Kata Kunci : Analisis Sentimen, *Logistic Regression*, Covid-19

RINGKASAN

Analisis Sentimen Pada Penggunaan *Hashtag* Covid-19 di Media Sosial *Twitter*,
Yuniar Fabi Putra, Nim E41171845, Tahun 2021, 56 hlm, Teknologi Informasi,
Politeknik Negeri Jember, Ery Setiyawan Jullev Atmadji, S.Kom, M.Cs.

Di era pandemik COVID – 19 ini banyak yang diharuskan untuk tetap dirumah saja. Jadi tidak menuntut kemungkinan bahwa penggunaan media sosial di Indonesia akan terus bertambah terlebih lagi penggunaan media sosial *Twitter* akan terus bertambah. Pertumbuhan ini membuat orang – orang menikmati berbagai kegiatan mereka di media sosial, termasuk informasi, keluhan dan lain – lain mengenai pandemik COVID – 19 ini. Di masa pandemik COVID – 19 ini, di Indonesia banyak pengguna memanfaatkan *Twitter* untuk memberikan pernyataan atau opini mengenai COVID – 19 ini dengan menggunakan *hashtag* COVID – 19 dan membuat sebuah *tweet* yang cenderung berisi opini positif maupun negatif. Kecenderungan *tweet* dengan *hashtag* COVID -19, dapat diketahui dengan analisa opini atau analisis sentimen. Analisis sentimen adalah upaya untuk melihat pendapat atau kecenderungan sentimen negatif atau positif berdasarkan teks *tweet* yang dibisa sesuai dengan topik yang dicari. Oleh karena itu dibutuhkan sebuah klasifikasi yang dapat menganalisis sentimen, terutama *tweet* yang berbahasa Indonesia. Penelitian dilakukan dengan metode klasifikasi menggunakan *Logistic Regression* dengan perhitungan pembobotan kata TF-IDF. Dengan menggunakan klasifikasi *Logistic Regression* mendapatkan akurasi score sebesar 74% dan hasil dari evaluasi sistem yaitu *presisi* 74,5%, *recall* 73%, dan *F-Measure* 73%.

PRAKATA

Puji syukur penulis panjatkan ke hadirat Allah Subhanahu Wa Ta'ala atas berkat rahmat dan karunia-Nya sehingga penulisan karya ilmiah berjudul “Analisis Setimen Pada Penggunaan *Hastag* Covid-19 di Media Sosial *Twitter*” dapat diselesaikan dengan baik.

Tulisan ini adalah laporan hasil penelitian yang dilaksanakan mulai bulan Juli 2021 bertempat di Politeknik Negeri Jember, sebagai salah satu syarat untuk memperoleh gelar Sarjana Sains Terapan (S.Tr.Kom) di Program Studi Teknik Informatika Jurusan Teknologi Informasi.

Penulis menyampaikan penghargaan dan ucapan terima kasih yang sebesar-besarnya sebagai berikut.

1. Saiful Anwar, S.Tp, MP selaku Direktur Politeknik Negeri Jember
2. Hendra Yufit Riskiawan, S.kom, M.Cs selaku Ketua Jurusan Teknologi Informasi
3. Trismayanti Dwi Puspitasari, S.Kom, M.Cs selaku Ketua Program Studi Teknik Informatika
4. Ery Setiyawan Jullev Atmadji, S.Kom, M.Cs selaku Dosen Pembimbing.
5. Seluruh dosen dan staf pengajar di Program Studi Teknik Informatika Politeknik Negeri Jember.
6. Teman-teman seangkatan di Teknik Informatika 2021, serta seluruh pihak yang turut membantu terlaksananya penelitian dan penulisan skripsi ini.

Skripsi ini masih kurang sempurna, mengharapkan kritik dan saran yang sifatnya membangun guna perbaikan di masa mendatang. Semoga tulisan ini bermanfaat.

Jember, Juli 2021

Penulis

DAFTAR ISI

HALAMAN SAMPUL.....	i
HALAMAN JUDUL.....	ii
HALAMAN PENGESAHAN.....	iii
HALAMAN PERNYATAAN	iv
HALAMAN PERSETUJUAN.....	v
MOTTO	vi
PERSEMBAHAN.....	vii
ABSTRACT.....	viii
ABSTRAK.....	ix
RINGKASAN	x
PRAKATA.....	xi
DAFTAR ISI.....	xii
DAFTAR GAMBAR	xiv
DAFTAR TABEL.....	xv
BAB 1. PENDAHULUAN	1
1. 1 Latar Belakang	1
1.2 Rumusan Masalah	2
1.3 Tujuan.....	2
1.4 Manfaat.....	3
BAB 2. TINJAUAN PUSTAKA	4
2.1 State Of The Art	4
2.2 Media Sosial	7
2.3 Twitter	8
2.3.1 Twitter API.....	10
2.4 Informasi	10
2.5 Penyebaran Informasi Kesehatan	11
2.6 Text Mining.....	12
2.7 Tahapan dalam <i>Text Mining</i>	13
2.8 Analisis Sentimen.....	15

2.9 <i>Logistic Regression</i> (Regresi Logistik)	16
BAB 3. METODE PENELITIAN.....	19
3.1 Tempat dan Waktu Penelitian	19
3.2 Alat dan Bahan	19
3.3 Metode Penelitian.....	20
BAB 4. HASIL DAN PEMBAHASAN.....	18
4.1 <i>Crawling</i> Data Twitter.....	18
4.2 Pelabelan Data	26
4.3 <i>Pre-procesing</i> Teks	26
4.4 Pembobotan Kata	29
4.5 Klasifikasi Dengan <i>Logistic Regression</i>	32
4.6 Evaluasi Sistem	35
4.7 Visualisasi Data	37
BAB 5. KESIMPULAN DAN SARAN	22
5.1 Kesimpulan.....	22
5.2 Saran.....	22
DAFTAR PUSTAKA	42
LAMPIRAN.....	44

DAFTAR GAMBAR

Gambar 2.1 Tahapan Text Mining.....	13
Gambar 3 .1 Tahapan Metode Penelitian.....	20
Gambar 4.1 Data Hasil Crawling.....	25
Gambar 4.2 Sourcecode Tokenisasi dan Steming.....	29
Gambar 4.3 Data Grafik Label.....	32
Gambar 4.4 Sourcecode Vectorisasi Tahap 1	33
Gambar 4.5 Hasil Vectorisasi Tahap 1	34
Gambar 4.6 Sourcecode Vectorisasi Tahap 2	34
Gambar 4.7 Hasil Vectorisasi Tahap 2	34
Gambar 4.8 Sourcecode Klasifikasi dan Akurasi	35
Gambar 4.9 Wordcloud Berlabel Negatif	38
Gambar 4.10 Wordcloud Berlabel Positif.....	38

DAFTAR TABEL

Tabel 2.1 Studi Literatur	6
Tabel 4.1 Tabel Pelabelan Data	26
Tabel 4.2 Tabel Tahapan Cleansing.....	27
Tabel 4.3 Tabel Tahapan Case Folding	27
Tabel 4.4 Tabel Tahapan Stopword	28
Tabel 4.5 Tabel Tahapan Tokenisasi dan Stemming.....	29
Tabel 4.6 Tabel Data.....	30
Tabel 4.7 Tabel Term Frequency	30
Tabel 4.8 Tabel DF (Document Frequency)	30
Tabel 4.9 Tabel IDF	31
Tabel 4.10 Confusion Matrix	35

BAB 1. PENDAHULUAN

1. 1 Latar Belakang

Media sosial di Indonesia sudah berkembang dengan pesat. Salah satu media sosial yang sering dipakai adalah Twitter. Twitter merupakan media sosial microblog yang memungkinkan pengguna untuk mengirimkan pesan yang dibatasi hingga 280 karakter per pesan yang biasa disebut tweet, dikirim oleh pengguna kepada pembacanya atau follower. Twitter dapat diakses melalui website atau aplikasi gratis yang bisa langsung di download melalui ponsel masing - masing pengguna. Twitter digunakan oleh semua orang untuk melakukan penilaian dan mengeluarkan opini mengenai segala sesuatu, melakukan posting dan rating dengan opini yang berbeda - beda. Dalam sistem Twitter, tanda # atau hashtag menunjukkan topik-topik khusus yang sedang dibahas. Fungsi hashtag dalam Twitter antara lain sebagai media pencarian dan menampilkan informasi lebih mudah, dan sebagai penanda topik yang sedang ramai atau trend.

Twitter adalah salah satu media sosial yang populer dan banyak digunakan pada saat ini. Twitter menempati peringkat kedua sebagai media sosial teraktif di Indonesia. Menurut laporan terbaru *We Are Social*, pada tahun 2020 disebutkan bahwa ada 175,4 juta pengguna internet di Indonesia. Dibandingkan tahun sebelumnya ada kenaikan 17 % atau 25 juta pengguna internet di Indonesia. Berdasarkan total populasi di Indonesia yang berjumlah 272,1 juta jiwa, maka artinya 64% setengah penduduk Indonesia telah merasakan akses ke dunia maya. Riset per November 2019, juga dari *We Are Social* menyebutkan pengguna Twitter di Indonesia mencapai 78 juta pengguna. Ditambah lagi dengan adanya era pandemik COVID – 19 ini banyak yang diharuskan untuk tetap dirumah saja. Jadi tidak menuntut kemungkinan bahwa penggunaan media sosial di Indonesia akan terus bertambah terlebih lagi penggunaan media sosial Twitter akan terus bertambah. Pertumbuhan ini membuat orang – orang menikmati berbagai kegiatan mereka di media sosial, termasuk informasi, keluhan dan lain – lain mengenai pandemik COVID – 19 ini.

Di masa pandemik COVID – 19 ini, di Indonesia banyak pengguna memanfaatkan Twitter untuk memberikan pernyataan atau opini mengenai COVID – 19 ini dengan menggunakan hastag COVID – 19 dan membuat sebuah tweet yang cenderung berisi opini positif maupun negatif. Kecenderungan tweet dengan hastag COVID - 19, dapat diketahui dengan analisa opini atau analisis sentimen. Analisis sentimen adalah upaya untuk melihat pendapat atau kecenderungan sentimen negatif atau positif berdasarkan teks tweet yang dibaca sesuai dengan topik yang dicari. Oleh karena itu dibutuhkan sebuah klasifikasi yang dapat menganalisis sentimen, terutama tweet yang berbahasa Indonesia.

Penelitian ini sebelumnya telah dilakukan oleh Aloysius Kurniawan Santoso, Astrid Noviriandini, Aliyah Kurniasih, Bagus Dwi Wicaksono, Ahmad Nuryanto tahun 2021 dengan judul “Klasifikasi Persepsi Pengguna Twitter Terhadap Kasus Covid-19 Menggunakan Metode *Logistic Regression*” . Pada penelitian tersebut menggunakan metode *Logistic Regression* dengan memberi variasi *hyperparameter* L2 dan *None*. Pada *hyperparameter* L2 diperoleh nilai akurasi 77% dan F1 score sebesar 74%. Dan variasi *hyperparameter* *None* diperoleh nilai akurasi 74% dan F1 score 70%.

Pada tugas akhir ini, peneliti akan lebih fokus pada analisis sentimen di media sosial twitter terhadap hastag COVID – 19 dengan menggunakan metode Logistic Regression.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang sudah diuraikan di atas, maka dapat diambil beberapa masalah yang dapat dibahas dalam tugas akhir ini yaitu :

1. Bagaimana menerapkan metode Logistic Regression untuk pengklasifikasian analisis sentimen terhadap penggunaan hastag COVID – 19 di Twitter untuk mengetahui sentimen positif atau negatif ?
2. Bagaimana perhitungan Logistic Regression untuk mengetahui tingkat akurasi yang dihasilkan metode tersebut dari data yang sudah ada ?

1.3 Tujuan

Adapun tujuan penelitian pada penelitian ini adalah sebagai berikut :

1. Untuk mengetahui penerapan metode Logistic Regression yang digunakan untuk klasifikasi sentimen terhadap penggunaan hastag COVID – 19.
2. Untuk mengetahui tingkat akurasi perhitungan Logistic Regression terhadap data yang sudah ada.

1.4 Manfaat

Adapun manfaat pada penelitian ini adalah sebagai berikut :

1. Bermaanfaat untuk memberikan informasi mengenai sentimen positif dan negatif terhadap penggunaan hastag COVID – 19 di masa pandemik ini.
2. Bermaanfaat bagi penulis untuk menambah wawasan ilmu baru dipelajari, sehingga dapat mengukur tingkat akurasi dari metode Logistic Regression.

BAB 2. TINJAUAN PUSTAKA

2.1 State Of The Art

Pada penelitian yang dilakukan, penulis mengacu pada studi literatur atau penelitian sebelumnya. Berikut ini penjabaran dari penelitian sebelumnya :

1. KLASIFIKAS PERSEPSI PENGGUNA TWITTER TERHADAP KASUS COVID-19 MENGGUNAKAN METODE *LOGISTIC REGRESSION* (Aloysius Kurniawan Santoso, Astrid Noviriandini, Aliyah Kurniasih, Bagus Dwi Wicaksono, Ahmad Nuryanto tahun 2021)

Pada hal ini peneliti mengelompokkan data menjadi 5 label diantaranya, ekstrim positif, positif, netral, negatif dan ekstrim negatif. Dalam penelitian ini menguji menggunakan metode *Logistic Regression* dengan memberi variasi *hyperparameter* L2 dan *None*. Pada *hyperparameter* L2 diperoleh nilai akurasi 77% dan F1 score sebesar 74%. Dan variasi *hyperparameter* *None* diperoleh nilai akurasi 74% dan F1 score 70%.

2. ANALISIS KEPRIBADIAN MELALUI TWITTER MENGGUNAKAN METODE *LOGISTIC REGRESSION* DENGAN PEMBOBOTAN TF-IDF DAN AHP (Kartika Prameswari, Erwin Budi Setiawan tahun 2019)

Pembobotan TF-IDF dan AHP ini dilakukan untuk memnentukan bobot disetiap fitur yang ada pada akun pengguna. Pada pendekatan linguistik digunakan pembobotan TF-IDF, sedangkan untuk pendekatan perilaku sosial menggunakan pembobotan AHP. Hasil dari klasifikasi ini dengan kedua pembobotan tersebut pada pendekatan perilaku sosial memiliki rata-rata akurasi sebesar 24,95%, sedangkan akurasi tertinggi pada pendekatan linguistik dengan ada pembagian data set 90 : 10 adalah 33,5 %.

3. PREDIKSI *BIG FIVE PERSONALITY* DENGAN *TERM FREQUENCY INVERSE DOCUMENT FREQUENCY* (TF-IDF) MENGGUNAKAN METODE *LOGISTIC REGRESSION* PADA PENGGUNA TWITTER (Rendo Zenico, Erwin Budi Setiawan, Fida Nurmala Nugraha tahun 2019)

Penelitian yang dilakukan menggunakan dua pendekatan, yaitu pendekatan linguistik dan pendekatan perilaku sosial dengan digunakannya fitur dari Twitter. Data yang digunakan adalah data dari 143 pengguna twitter dengan

perbandingan data 70:30. Menggunakan pembobotan *Term Frequency Inverse Document Frequency* (TF-IDF) dan *Logistic Regression* sebagai algoritma klasifikasi, akurasi yang dihasilkan oleh sistem yang dibangun pada tugas akhir ini 69% untuk pendekatan perilaku sosial dan 76,20% untuk pendekatan linguistik dan pendekatan perilaku sosial.

Tabel 2.1 Studi Literatur

NO	Nama Peneliti	Judul	Hasil Penelitian
1.	Aloysius Kurniawan Santoso, Astrid Noviriandini, Aliyah Kurniasih, Bagus Dwi Wicaksono, Ahmad Nuryanto tahun 2021	Klasifikas Persepsi Pengguna Twitter Terhadap Kasus Covid-19 Menggunakan Metode <i>Logistic Regression</i>	Metode : <i>Logistic Regression</i> Hasil : Pengujian dengan hyperparamater L2 merupakan pengujian yang dapat menghasilkan model <i>logistic regression</i> yang lebih baik dalam menentukan klasifikasi terkait komentar pada twitter mengenai virus corona atau Covid-19 masuk dalam kategori positif, negatif, atau netral
2.	Rendo Zenico, Erwin Budi Setiawan, Fida Nurmala Nugraha tahun 2019	Analisis Kepribadian Melalui Twitter Menggunakan Metode <i>Logistic Regression</i> Dengan Pembobotan TF-IDF Dan AHP	Metode : <i>Logistic Regression</i> Hasil : Akurasi terbaik dari pendekatan linguistik melalui pembobotan TF-IDF sebesar 33,55% dengan pembagian data set data uji dan dat latih 90:10. Dengan pembobotan AHP didapatkan akurasi terbaik dengan fitur yang digunakan sebanyak 15 fitur yaitu <i>followers</i> , <i>following</i> , jumlah <i>tweet</i> , <i>like</i> , <i>emoji</i> , panjang bio, <i>website</i> , media URL, <i>retweet</i> , <i>hashtag</i> , panjang <i>tweet</i> , mean kata, dan mentio dengan akuarasi sebesar 40,1%. Bobot yang paling berpengaruh adalah dalam pembobotan AHP yaitu <i>followers</i>

3.	Rendo Zenico, Erwin Budi Setiawan, Fida Nurmala Nugraha tahun 2019	Prediksi <i>Big Five</i> <i>Personality</i> Dengan <i>Term Frequency</i> <i>Inverse Document</i> <i>Frequency</i> (Tf-Idf) Menggunakan Metode <i>Logistic Regression</i> Pada Pengguna Twitter	Metode : <i>Logistic Regression</i> Hasil : Pengujian pendekatan perilaku sosial dengan atribut Twitter, akurasi yang terbaik yang dihasilkan 69 %. Fitur atribut yang berpengaruh <i>follower</i> , <i>following</i> , <i>media_url</i> , <i>url</i> , <i>rata2_karakter</i> , <i>retweet</i> , <i>kata</i> , <i>rata_kata</i> , <i>karakter</i> , <i>mention</i> , <i>tanda_baca</i> , <i>huruf_besar</i> , <i>hashtag</i> . Pengujian pendekatan linguistik denga TF-IDF dan perilaku sosial, menghasilkan akurasi yang lebih baik. Akurasi yang didapat dari sekenario ini adalah 76,20%. Penggunaan bentuk kata unigram pada TF-IDF juga belum mampu untuk memaksimalkan uji similaritas data.
----	---	--	---

2.2 Media Sosial

Media Sosial adalah sebuah media daring, dengan para penggunanya bisa dengan mudah berpartisipasi, berbagi, dan menciptakan isi *blog*, jejaring sosial, wiki, forum dan dunia virtual. *Blog*, jejaring sosial, dan wiki merupakan bentuk media sosial yang paling umum digunakan oleh masyarakat di seluruh dunia. Andreas Kaplan dan Michael Haenlein mendefinisikan media sosial sebagai “sebuah kelompok aplikasi berbasis internet yang dibangun di atas dasar ideologi dan teknologi Web 2.0 dan memungkinkan penciptaan dan pertukaran user – generated content”.

2.3 Twitter

Twitter merupakan layanan jejaring sosial dan mikroblog daring yang memungkinkan penggunanya untuk mengirim dan membaca pesan berbasis teks hingga 140 karakter akan tetapi pada tanggal 07 November 2017 bertambah hingga 280 karakter yang dikenal dengan sebutan kicauan (*tweet*). Twitter didirikan pada bulan maret 2006 oleh Jack Dorsey, dan situs jejaring sosialnya diluncurkan pada bulan Juli. Di Twitter, pengguna tak terdaftar hanya bisa membaca kicauan, sedangkan pengguna terdaftar bisa menulis kicauan melalui antarmuka situs web, pesan singkat (SMS), atau melalui berbagai aplikasi untuk perangkat seluler.

Twitter mengalami pertumbuhan yang pesat dan dengan cepat meraih popularitas di seluruh dunia. Hingga bulan Januari 2013, terdapat lebih dari 500 juta pengguna terdaftar di Twitter, 200 juta diantaranya adalah pengguna aktif. Lonjakan pengguna Twitter umumnya berlangsung saat terjadinya peristiwa – peristiwa populer. Pada awal 2013, pengguna Twitter mengirimkan lebih dari 500 juta kicauan per hari, dan Twitter menangani lebih dari 1,6 miliar permintaan pencarian per hari. Hal ini menyebabkan posisi Twitter naik ke peringkat kedua sebagai situs jejaring sosial yang paling sering dikunjungi di dunia, dari yang sebelumnya menempati peringkat dua puluh dua. Tingginya popularitas Twitter menyebabkan layanan ini telah dimanfaatkan untuk berbagai keperluan dalam berbagai aspek, misalnya sebagai sarana protes, kampanye politik, sarana pembelajaran, dan sebagai media komunikasi darurat.

Pengguna media dapat menggunakan *Twitter* sebagai sarana untuk menciptakan konten media dengan memanfaatkan fitur-fitur yang berada didalamnya. Menurut Brian J. Dixon fitur yang terdapat dalam *Twitter* sebagai berikut :

a. *Followers* dan *Following*

Followers (pengikut) merupakan akun atau orang yang mengikuti akun yang lain, sedangkan *Following* (mengikuti) merupakan akun atau orang yang diikuti akun yang lain.

b. *Direct Message*

Twitter juga memungkinkan untuk mengirim pesan pribadi ke pengguna yang mengikuti akun tersebut. Pada dasarnya adalah program e-mail yang diterapkan ke *Twitter*. Jika pengguna *Twitter*.

c. *Twitter Search*

Fitur ini merupakan fitur paling kuat dari *Twitter*. Dalam fitur ini memberi kemudahan pengguna untuk mencari orang-orang tertentu, kata kunci, subjek, dan tempat-tempat.

d. *Trending Topics*

Salah satu bagian yang paling menarik dari *Twitter* adalah *trending topics*. Merupakan topik yang sering disebut atau dibicarakan di *Twitter* pada waktu tertentu yang berkisar dari berita, olahraga, dan barang-barang hiburan yang menghibur.

e. *Latest News*

Twitter memungkinkan penggunaanya dengan cepat mengejar ketinggalan berita terbaru. Begitu seorang tahu tentang cerita terbaru, pengguna dapat memposting informasi tersebut di *Twitter*, dan dalam beberapa detik konten yang dibagi muncul di internet. (Dixon, 2012: 43-45)

Jenis media sosial *Twitter* dapat menciptakan serta menggerakkan komunitas, mengendalikan *traffic* di blog, atau website. Sedangkan untuk kekuatan *Twitter* menurut Puntoadi (2011 : 129-131) adalah :

- *Following* : *Twitter* dapat mengidentifikasi jumlah akun yang diikuti.
- *Follower* : *Twitter* dapat mengidentifikasi jumlah akun yang mengikuti.
- *Updates (Tweet)* : *Twitter* mendeteksi seberapa sering orang melakukan posting.

- *Retweet* : dengan me-*retweet* status akun lain menunjukkan bahwa status atau *tweet* tersebut menarik. *Tweet* tersebut dapat dibaca oleh seluruh *Follower*.
- *Name-tangging* : beberapa orang akan terlibat dalam pembicaraan apabila mencantumkan username suatu akun lain setiap status yang ditulis.
- Kecepatan : konsistensi dan frekuensi diperlukan untuk mengimbangi kecepatan perubahan informasi yang beredar di *Twitter*.

2.3.1 Twitter API

Twitter Application Programming Interface (API) menyediakan *Streaming* API dan dua diskrit REST API. Pengguna bisa menbisakan akses *tweet* di twitter kemudian dikumpulkan dan difilter sesuai kebutuhan. API yang berbasis permintaan HTTP, GET, POST, dan DELETE bisa dipakai untuk mengakses data (Albert Bifet, 2010)

Dalam terminologi *twitter*, setiap pesan menggambarkan status seorang pengguna. Berdasarkan API *streaming* para pengguna bisa mengakses status publik hampir secara *realtime*, termasuk balasan dan *mention* yang dibuat oleh akun – akun publik, status yang dibuat oleh akun – akun terproteksi juga pesan yang tak bisa diakses. Salah satu properti menarik dari API *streaming* adalah bisa memfilter status dengan memakai metrik yang seringkali adalah status repitisi, dll. API tersebut memakai autentikasi HTTP dan membutuhkan akun *Twitter* yang valid dan data bisa diterima sebagai XML.

2.4 Informasi

Informasi adalah jumlah ketidak pastian yang dapat diukur dengan cara mereduksi sejumlah alternatif pilihan yang tersedia. Informasi berkaitan dengan situasi yang tidak pasti. Semakin tidak pasti suatu situasi, maka semakin banyak pula alternatif yang dapat digunakan secara berturut-turut dan tumpang tindih (reduktif) untuk mengurangi ketidak pastian tertentu. Teori tersebut di jelaskan oleh

Sendjaja (1998) dalam Kriyantono (2009 : 380) yang menjelaskan konsep dasar teori informasi yang berasal dari Claude Shannon dan Warren Weaver dalam buku *The Mathematical Theory of Communication*.

Informasi adalah pengukuran ketidakpastian atau entropi dalam suatu situasi. Semakin besar ketidakpastian maka semakin banyak informasi yang dibutuhkan. Bila situasi dapat diperkirakan seluruhnya, maka tidak ada informasi tersaji. Kondisi ini disebut juga dengan istilah negentropi, dengan kata lain, suatu situasi dengan mana seluruhnya kita kenal, berarti tidak memiliki informasi baru bagi kita (Littlejohn (1998) dalam Kriyantono, 2009 : 380-381).

2.5 Penyebaran Informasi Kesehatan

Media sosial telah menjadi sesuatu yang memungkinkan dari penyebaran informasi, kolaborasi dan koordinasi untuk beragam hal mulai dari hal pribadi hingga masalah dibidang kesehatan. Terlebih dimasa pandemik ini media sosial menjadi media paling besar dalam penyebaran informasi mengenai pandemik ini.

Media sosial sekarang menyediakan ruang untuk mendiskusikan kondisi medis di luar kantor penyedia layanan kesehatan. Pasien dan keluarga mereka menggunakan teknologi media sosial untuk berbagai pengalaman dan temuan mereka dan mendidik orang lain dengan kondisi yang sama. Mereka mengemas ulang informasi yang mereka temukan untuk orang lain, menciptakan forum-forum untuk penemuan pengetahuan lain dan diskusi.

Menurut Lapointe; dkk (2013), media sosial menyediakan forum untuk melaporkan pengalaman pribadi, mengajukan pertanyaan, dan menerima umpan balik langsung dengan sesama pengguna. Pekerja profesional kesehatan juga menggunakan alat-alat media sosial sebagai platform penyebaran informasi dibidang kesehatan. Hal ini menunjukkan bahwa alat media sosial telah memungkinkan kolaborasi antara individu. Hal ini dapat terjadi dalam konteks seperti karyawan yang bekerja bersama-sama dalam batas organisasi formal, untuk konteks di mana individu tersebar dapat

terhubung satu sama lain karena ketersediaan media sosial (Lapointe; dkk, 2013:2).

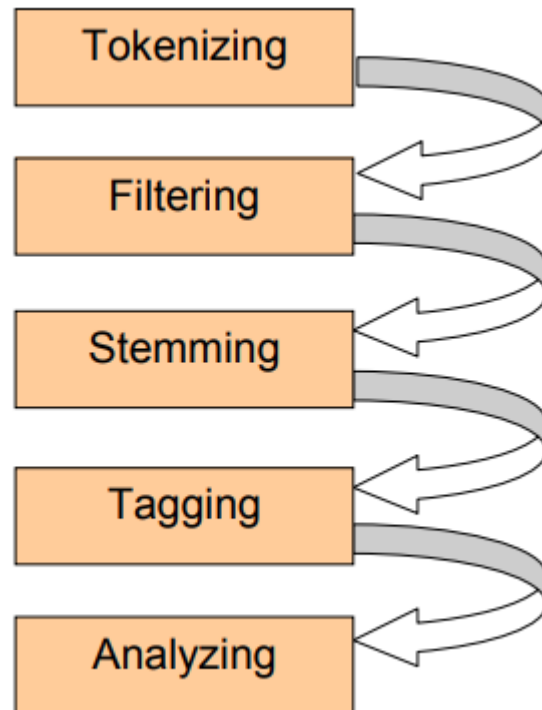
2.6 Text Mining

Text mining, menurut Milkha Harlin (2006) yaitu *text mining* mempunyai definisi penambangan data berupa teks, di mana sumber data biasanya dibiasakan dari dokumen dan bertujuan untuk mencari kata-kata yang bisa mewakili isi dari dokumen sehingga bisa dilakukan analisis keterhubungan antar dokumen. Proses ini bertujuan untuk menggabungkan informasi yang berhasil diekstraksi dari berbagai sumber (Hearst, 2003). Dengan *text mining* tugas-tugas yang berhubungan dengan penganalisaan teks dengan jumlah yang besar, penemuan pola serta penggalian informasi yang mungkin berguna dari suatu teks bisa dilakukan.

Text mining atau penambangan teks merupakan proses ekstraksi pola berupa informasi dan pengetahuan yang berguna dari sejumlah besar sumber data teks. Jenis masukan untuk penambangan teks ini disebut data tak terstruktur dan adalah pembeda utama dengan penambangan data yang memakai data terstruktur atau basis data sebagai masukan. *Text mining* adalah proses yang terdiri atas dua tahap yaitu diawali dengan menerapkan struktur terhadap sumber data teks dan dilanjutkan dengan mengekstraksi informasi dan pengetahuan yang relevan dari data teks terstruktur tersebut dengan memakai teknik dan alat yang sama dengan data mining.

2.7 Tahapan dalam *Text Mining*

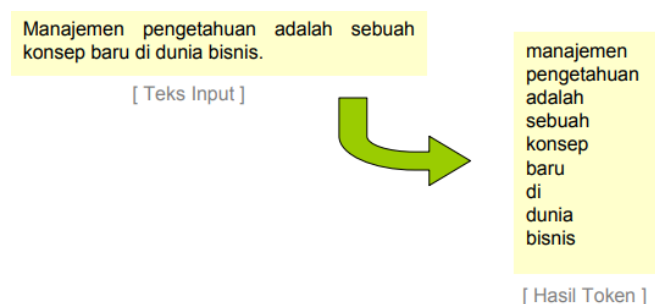
Tahapan yang dilakukan secara umum adalah :



Gambar 2.1 Tahapan *Text Mining*

a. *Tokenizing*

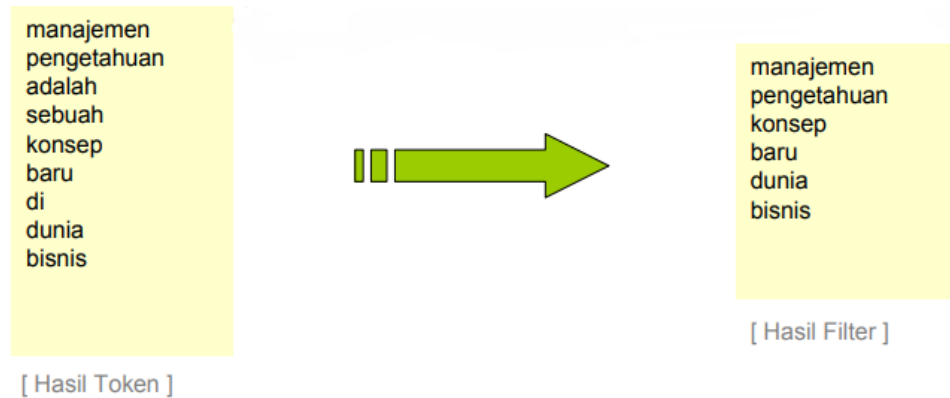
Tahap *tokenizing* adalah tahap pemotongan string input berdasarkan tiap kata yang menyusunnya. Contoh dari tahap ini adalah sebagai berikut :



Gambar 2.1 Tahapan *Tokenizing*

b. *Filtering*

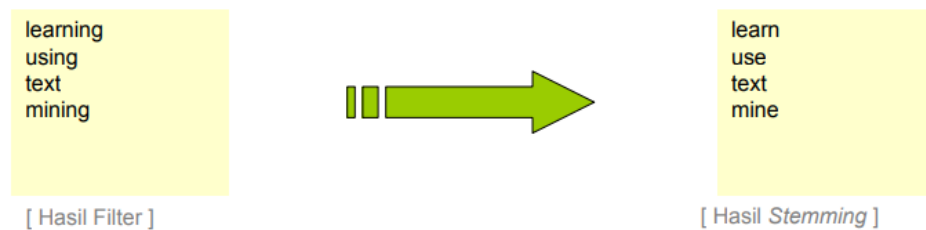
Filtering adalah tahap mengambil kata-kata penting dari hasil token. Bisa menggunakan algoritma *stoplist* (membuang kata yang kurang penting) atau *word list* (menyimpan kata penting). Contoh dari tahap ini adalah sebagai berikut :



Gambar 2.2 Tahapan Filtering

c. *Stemming*

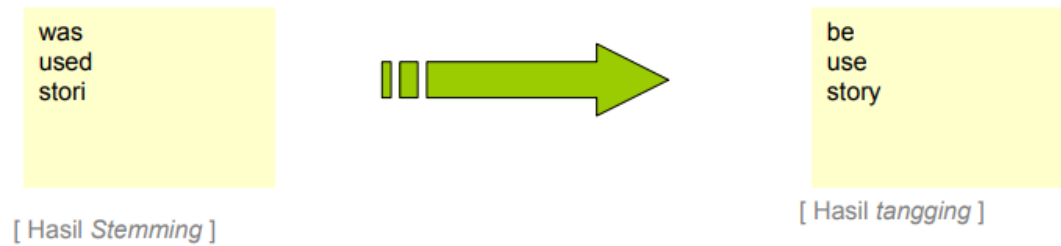
Pada tahapan ini adalah mencari root kata dari tiap kata hasil *filtering*. Contoh dari tahap ini adalah sebagai berikut :



Gambar 2.3 Tahapan Stemming

d. *Tagging*

Tahap *Tagging* adalah tahap mencari bentuk awal atau *root* dari tiap kata lampau atau kata hasil *stemming*. Contoh dari tahap ini sebagai berikut :



Gambar 2.4 Tahapan *Tagging*

e. *Analyzing*

Tahap *analysing* merupakan tahap penentuan seberapa jauh keterhubungan antar kata-kata antar dokumen yang ada.

2.8 Analisis Sentimen

Analisis sentimen adalah bidang studi analisis pendapat atau opini, sentimen, evaluasi, penilaian, sikap, dan emosi seseorang terhadap suatu barang, organisasi, orang, masalah yang konkrit, serta peristiwa. Kata “analisis sentimen” pertama muncul tahun 2003 oleh Nasukawa dan Yi, sementara “*opinion mining*” muncul tahun 2003 oleh Dave. Lawrence dan Pennock. (Liu, 2012).

Analisis sentimen juga merupakan metode yang digunakan untuk mengekstrak data opini, memahami serta mengolah tekstual data secara otomatis untuk melihat sentimen yang terkandung dalam sebuah opini. Sebuah metode yang berbeda untuk menentukan sentimen merupakan pemakaian sistem skala di mana kata – kata umumnya terkait mempunyai sentimen negatif, netral atau positif dengan mereka beri nomor pada skala -5 sampai +5 (paling negatif hingga yang paling positif) dan ketika sepotong teks terstruktur dianalisis dengan pemrosesan bahasa alami, konsep selanjutnya dianalisis untuk memahami kata-kata ini dan bagaimana mereka berhubungan dengan konsep. Setiap konsep kemudian diberi skor berdasarkan bagaimana kata-kata sentimen berhubungan dengan konsep, dan skor yang terkait. Hal ini memungkinkan gerakan untuk pemahaman yang lebih canggih dari sentimen berdasarkan skala 11 titik.

Secara umum, analisis sentimen ini dibagi menjadi 2 kategori umum (Schneider, 2005) :

a. *Coarse-grained sentiment analysis*

Kategori ini melakukan proses analisis pada level dokumen. Jadi kita mencoba mengklasifikasikan orientasi sebuah dokumen secara keseluruhan. Orientasi ini ada 3 kategori : positif, netral, negatif. Akan tetapi , ada juga yang menjadikan nilai orientasi ini bersifat kontinu atau tidak diskrit.

b. *Fined-grained sentiment analysis*

Kategori ini yang sedang marak sekarang. Maksudnya adalah para peneliti sebagian besar fokus pada jenis ini. Obyek yang ingin diklasifikasi bukan pada dokumen melainkan pada sebuah kalimat pada suatu dokumen.

Analisis sentimen terdiri dari 3 subproses besar (Eyheramendy, Lewis, & Madigan, 2003). Masing-masing subproses ini bisa kita jadikan bahan atau topik riset secara terpisah yaitu :

1. *Subjectivity Classification* , menentukan kalimat yang merupakan opini.
2. *Orientation Detection*, setelah berhasil diklasifikasikan untuk kategori opini, sekarang kita tentukan positif, negatif, netral.
3. *Opinion Holder and Target Detection*, menentukan bagian yang merupakan Opinion Holder dan bagian yang merupakan target.

Penelitian ini berfokus membahas mengenai *Orientation Detection* terhadap suatu kalimat dalam opini, yaitu bagaimana menentukan opini positif dan negatif, maupun netral dalam sebuah *tweet*.

2.9 Logistic Regression (Regresi Logistik)

Regresi Logistik merupakan sebuah metode statistik yang diterapkan untuk memodelkan variabel respon yang bersifat kategori (skala nominal/ordinal) berdasarkan satu atau lebih pengubah prediktor yang dapat berupa variabel kategori maupun kontinu (skala interval atau rasio).

Regresi Logistik bagian dari analisis regresi yang digunakan jika variabel dependent (respon) merupakan variabel dikotomi. Variabel dikotomi biasanya hanya terdiri atas dua nilai, yang kemunculannya mewakili atau tidak adanya suatu kejadian yang biasanya diberi angka 0 atau 1 (Nirwana, 2015).

Tidak seperti regresi linier biasa, regresi logistik tidak mengasumsikan dari hubungan antara variabel independent dan dependent secara linier. Regresi logistik adalah regresi non linier dimana model yang ditentukan akan mengikuti pola kurva linier. Pada regresi logistik membentuk variabel prediktor yang merupakan kombinasi linier dari variabel independent. Nilai dari variabel prediktor ini kemudian ditransformasikan menjadi probabilitas dengan fungsi logit.

Regresi logistik bertujuan untuk menanggulangi kelemahan dari LPM (Linier Probability Model) yang dapat memberi hasil kurang memuaskan, karena menghasilkan probabilitas taksiran yang kurang dari nol atau lebih dari satu. Dalam hal ini, yang mampu menjamin nilai variabel dependent terletak antara 0 dan 1 sesuai dengan teori probabilitas adalah dengan model CDF (*Cumulative Distribution Function*). Dengan CDF yang memiliki dua sifat yaitu: 1) jika variabel bebas naik, maka juga ikut naik, tetapi tidak pernah melewati rentangan 0 – 1, dan 2) hubungan antara dan adalah non linear, sehingga, tingkat perubahannya tidak sama, tingginya semakin besar kemudian mengecil. Ketika nilai probabilitasnya mendekati nol, tingkat penurunannya semakin kecil, demikian juga ketika nilai probabilitasnya mendekati satu, maka tingkat tingginya semakin kecil. Secara umum, persamaan regresi logistik untuk k variabel dependent (Nirwana, 2015). Terdapat pada persamaan (2.1).

$$\ln[\text{odds}(T/X_1, X_2, \dots, X_k)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k \quad (2.1)$$

Regresi Logistik membentuk variabel prediktor ($\ln (P/(1-P))$) merupakan kombinasi linier dari variabel independent. Dari nilai variabel prediktor kemudian ditransformasikan menjadi probabilitas dengan fungsi logit.

Jadi model regresi linier sederhana terdapat pada persamaan (2.2).

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (2.2)$$

Dimana Y_i merupakan variabel respon, β_0 dan β_1 merupakan parameter, ε_i merupakan galat ke i , di mana $i = 1, 2, \dots, n$. Apabila persamaan 2.2 merupakan model regresi yang tidak memiliki intersep maka persamaan tersebut terdapat pada persamaan (2.3).

$$Y_i = \beta_1 X_i + \varepsilon_i \quad (2.3)$$

Di mana Y_i dan X_i pengamatan yang dilakukan pada Y dan X dengan galat ε di mana $i = 1, 2, \dots, n$. Bila pengamatan diambil sebanyak n maka persamaan ini terdapat pada persamaan (2.4), (2.5), dan (2.6).

$$Y_1 = \beta_1 X_1 + \varepsilon_1 \quad (2.4)$$

$$Y_2 = \beta_1 X_2 + \varepsilon_2 \quad (2.5)$$

$$Y_n = \beta_n X_n + \varepsilon_n \quad (2.6)$$

Yang mana ε merupakan variable acak normal bebas dan β_1 adalah parameter dengan $E(\varepsilon) = 0$ dan $\text{var}(\varepsilon) = \sigma^2$. Variabel respon dalam persamaan regresi tidak hanya dipengaruhi oleh variabel bebas yang bersifat kuantitatif saja (seperti umur, pendapatan, harga dan sebagainya), tetapi seringkali juga dipengaruhi oleh variabel yang bersifat kualitatif (seperti jenis kelamin, musim, warna dan sebagainya). Jadi berdasarkan variabel-variabel yang bersifat kualitatif maka dapat diketahui regresi dengan variabel kualitatif yang hanya memiliki 2 nilai yaitu nilai 1 dan 0, 12 salah satu model yang memiliki variabel yang bersifat kualitatif yaitu model regresi logistik.

BAB 3. METODE PENELITIAN

3.1 Tempat dan Waktu Penelitian

3.1.1 Tempat Penelitian

Tempat penelitian yang digunakan dalam proses penelitian ini berada di rumah pribadi.

3.1.2 Waktu penelitian

Waktu penelitian yang dilakukan dalam proses penelitian ini yaitu selama 6 bulan.

3.2 Alat dan Bahan

3.2.1 Alat

Alat-alat yang digunakan dalam penelitian ini yaitu perangkat keras dan perangkat lunak :

a. Perangkat keras :

Laptop	: Asus X555BA RAM 4GB Processor AMD Dual Core A9-9420
Handphone	: Redmi 4A Android 7.1.2 N2G47H MIUI Global 10.2.3 RAM 2GB

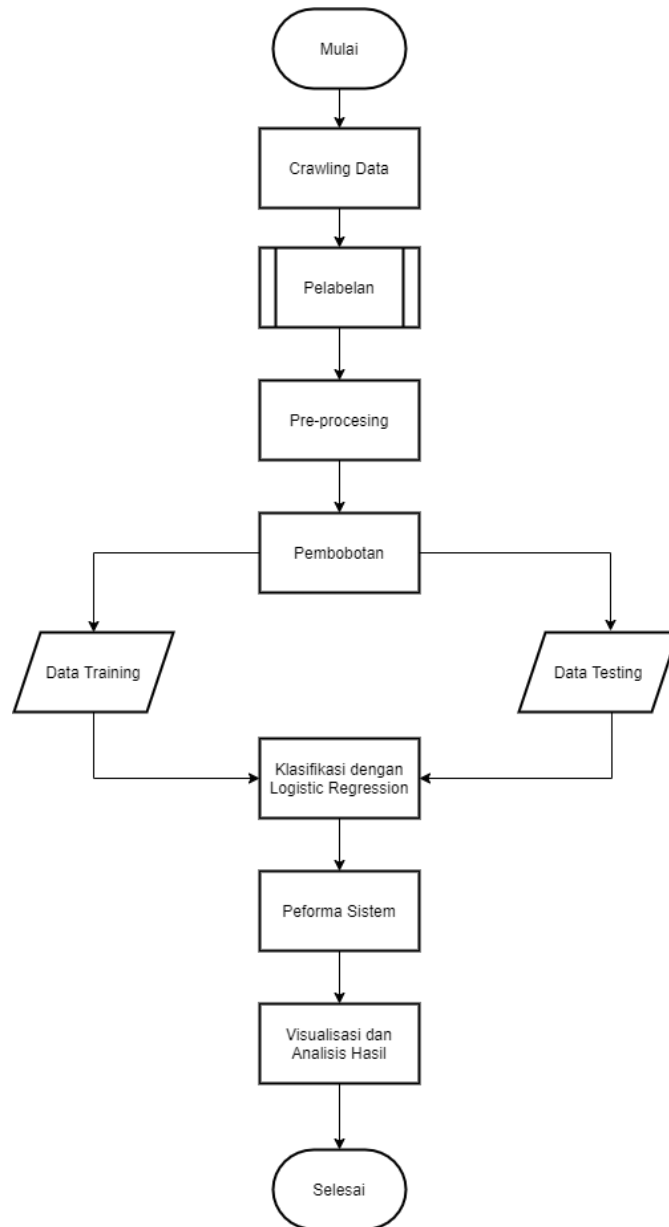
b. Perangkat lunak : Windows 10, Command Prompt, Software Jupyter Notebook

3.2.2 Bahan

Bahan yang digunakan dalam penelitian ini adalah trend hastag dari media sosial *twitter* hastag covid-19 . Bahan penelitian ini didapatkan melalui proses *crawling*.

3.3 Metode Penelitian

Dalam penelitian ini, digunakan metode gambar di bawah ini :



Gambar 3 .1 Tahapan Metode Penelitian

3.3.1 *Crawling Data*

Mendapatkan data yang dibutuhkan dalam pengambilan data penelitian ini dilakukan crawling data twitter. Pertama harus mempunyai atau sudah registrasi sebagai *developer* twitter untuk mendapatkan kode akses API twitter. Setelah

mendapatkan kode akses, langkah selanjutnya dengan menggunakan *software Jupyter Notebook* untuk melakukan crawling data twitter. *Crawling* dilakukan dengan menentukan kata kunci yang dicari dan menentukan jumlah data yang diambil dan disimpan dalam bentuk dokumen microsoft excel berekstensi .csv.

3.3.2 Labeling

Data yang sudah diperoleh dari proses *scraping* dilakukan labelisasi dengan dibuatkan sebuah kolom baru pada data untuk relevansi data dimana kolom tersebut diisi dengan nilai 0 atau 1 secara manual. Nilai 0 menandakan bahwa *tweet* bersifat negatif dengan hastag pada *tweet* tersebut, sedangkan nilai 1 menandakan yang sebaliknya dimana *tweet* bersifat positif dengan *hashtag* pada *tweet* tersebut.

3.3.3 Pre - procesing

Dalam tahap ini merupakan proses mengubah data menjadi terstruktur dan bisa digunakan sesuai dengan kebutuhan dalam penelitian. Pada penelitian ini, semua data *tweet* yang telah diambil dari proses *crawling* akan diproses sebagai berikut :

- a. *Case Folding* : Merubah semua huruf menjadi tidak kapital
- b. *Tokenizing* : Mengurangi dan kalimat tersebut menjadi kata, menghilangkan tanda baca, spasi dan karakter yang tidak diperlukan.
- c. *Filtering* : Memili kata penting dari hasil *tokenizing* dengan memanfaatkan stoplist atau kamus yang sudah didefinisikan.
- d. *Stemming* : Membentuk kata menjadi bentuk kata dasarnya.

Tabel 4.1 Ilustrasi *Preprocessing*

Kalimat	<i>Case Folding</i>	<i>Tokenizing</i>	<i>Filtering</i>	<i>Stemming</i>
Hari Senin	hari senin	hari senin	hari senin	hari senin
pasti	pasti menjadi	pasti	pasti	pasti jadi
menjadi	yang terbaik	menjadi	menjadi	baik
yang	!!	yang terbaik	terbaik	
TERBAIK !!				

3.3.5 Pembobotan Kata

Setelah dilakukan preproses , jadi setiap kata dalam dokumen diberi bobot dengan metode yaitu TF murni dan yang kedua yaitu TF-IDF. Dalam proses ini setiap kata yang ada pada dokumen tersebut mempunyai bobot sendiri – sendiri . Setelah proses tersebut selesai dilanjutkan dengan proses klasifikasi.

3.3.6 Klasifikasi dengan *Logistic Regression*

Data yang sudah dilanjutkan dengan pengujian klasifikasi. Pada proses ini, data akan dimasukkan kedalam pengujian perhitungan *Logistic Regression*. *Output* perhitungan ini akan menghasilkan yang performansinya akan dilakukan pengujian.

3.3.7 Model prediksi

Model prediksi merupakan sistem klasifikasi yang sudah dibuat dengan *Logistic Regression*. Model prediksi yang sebelumnya sudah dibangun akan diuji dengan data uji yang sudah disiapkan sebelumnya. Keluaran dari model prediksi ini akan menunjukkan nilai performansi sistem yang telah dibuat.

3.3.8 Perhitungan Peforma Sistem

Setelah data uji sudah diklasifikasi, maka dilakukan perhitungan *precision*, *recall*, dan akurasi akan dilakukan pada proses ini, yang bertujuan untuk mengukur performansi sistem.

3.3.9 Visualisasi dan Analisis Hasil

Setelah hasil klasifikasi sentimen terhadap COVID-19, langkah selanjutnya dilakukan visualisai dengan *wordcloud*. Kegunaan dari *wordcloud* tersebut adalah untuk mengetahui kata-kata mana yang sering muncul yang berpengaruh besar dalam pembuatan model klasifikasi sistem tersebut. Selain *wordcloud* juga divisualisasikan dalam bentuk grafik agak mudah dipahami.

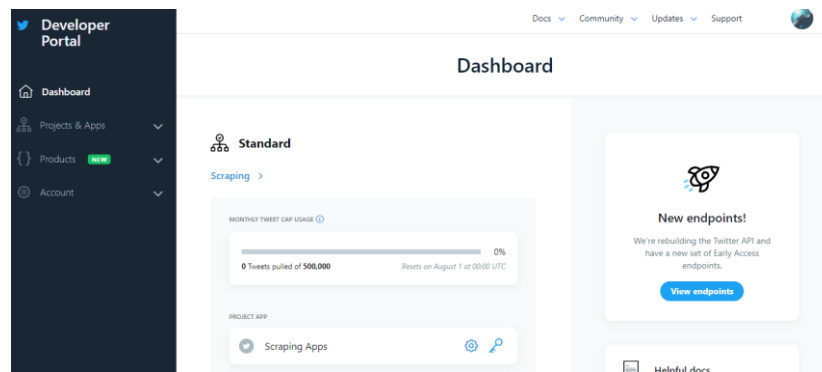
BAB 4. HASIL DAN PEMBAHASAN

4.1 *Crawling* Data Twitter

Langkah pertama yang harus dilakukan pada penelitian ini yaitu menarik data atau *crawling* data twitter. *Crawling* dilakukan secara berkala. Dikarenakan pada saat proses *crawling* terbatas pada tujuh hari terakhir dari waktu pencarian.

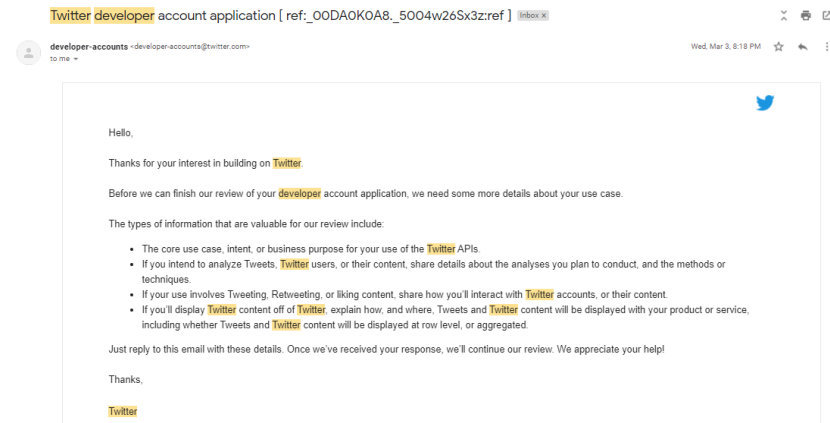
4.1.1 Pendaftaran Akun Developer Twitter

Untuk mendapatkan akses API twitter dilakukan pendaftaran akun dengan memakai akun twitter penulis sebagai akun untuk developer twitter. Pada tahapan ini melakukan registrasi app yang merupakan identifier penggunaan token API yang akan dimanfaatkan untuk mengambil atau menarik data pada twitter. App yang didaftarkan tidak dapat digunakan secara langsung, tetapi melalui tahap verifikasi oleh pidak developer twitter apakah tujuan dari penggunaan API tersebut dan tidak melanggar ketentuan.



Gambar 4.1 Pendaftaran Akun *developer* portal API

Tahap lanjutan setelah proses registrasi yaitu dilakukannya di website resmi di twitter developer. Adapun pertanyaan yang diajukan oleh pihak twitter yang berkaitan dengan tujuan atau studi kasus yang akan dilakukan dengan data yang sudah didapat dari twitter, dan pengajuan pertanyaan yang diajukan ini melalui media email penulis.



Gambar 4.2 Email pengajuan pertanyaan

4.1.2 Akses Token

Setelah diajukannya pertanyaan dan sudah disetujui pihak developer twitter, pengajuan app developer akan dilakukan review oleh pihak twitter dan apakah memenuhi syarat untuk diberikan akses ataupun ditolak. Setelah akun developer memenuhi syarat, pihak twitter akan menghubungi pemilik akun developer bahwa pengajuan telah disetujui.

4.1.3 Penarikan Data (*Crawling*)

Pada proses *crawling* ini menggunakan software *Jupyter Notebook*. Bahasa pemrograman yang digunakan adalah Python, dan di gunakan librari pada python yaitu *Tweepy* yang dalam proses nya dibutuhkan akses token yang sudah didapatkan sebelumnya untuk memulai penarikan data. *Crawling data* di twitter ini dilakukan kembali dikarenakan ditahun 2021 ada lonjakan kasus covid-19, maka dari itu dilakukan crawling ditanggal 3 Juli 2021. Dari proses *crawling* tersebut mendapatkan data yang didapatkan sebanyak 1000 data. Data tersebut merupakan data yang terdiri dari beberapa kelas atau kategori diantaranya positif, negatif, dan netral. Berikut adalah Gambar 4.1 merupakan contoh dari data yang sudah diambil.

	username	tweetcreatedts	text
0	SHWBAPAK	2021-07-02 23:59:53	Ivermectin telah disimpulkan tidak bisa diguna...
1	k_wandani	2021-07-02 23:59:53	Halo, yang punya info kamar RS kosong boleh ka...
2	sa_ae16	2021-07-02 23:59:51	Selamat jalan, maestro. Covid-19 merenggut san...
3	ayuyayuyaya	2021-07-02 23:59:50	Pemerintah secara resmi memberlakukan PPKM Dar...
4	3_jotr	2021-07-02 23:59:49	Selamat jalan, maestro. Covid-19 merenggut san...
5	RameshRaoAKS	2021-07-02 23:59:48	Data terbaru Johns Hopkins University menunj...
6	PutraWadapi	2021-07-02 23:59:48	Alhamdulillah, Madrasah Aliyah Negeri (MAN) 7 ...
7	BeritaKodim	2021-07-02 23:59:47	Babinsa Koramil 12/Pundong Kodim 0729/Bantul S...
8	SopoJarwo	2021-07-02 23:59:46	Anggara Bansos Covid-19 triliunan rp masih min...
9	ninalistya	2021-07-02 23:59:44	pasien meninggal siang ini di IGD dx/ Gagal na...
10	weirdianae	2021-07-02 23:59:44	KKM sendiri akui 80% kluster di Selangor adala...
11	nrrflr	2021-07-02 23:59:42	Berdasarkan rekomendasi WHO, tes untuk Covid-1...
12	Junita27468818	2021-07-02 23:59:42	Ndeer..tulang jangan dulu ada acara makan bers...
13	RameshRaoAKS	2021-07-02 23:59:36	Pahang catat kadar keboleangkitan tertinggi ...
14	KikaOyes	2021-07-02 23:59:34	Baru Pulang dr LN Langsung Rapat di DPR lnAngg...
15	KWKew3	2021-07-02 23:59:30	Kadar keboleangkitan Covid-19 atau Ro/Rt pad...
16	alzforall	2021-07-02 23:59:21	Urusan Covid-19 diserahkan kepada pebisnis tuk...
17	nadadjul	2021-07-02 23:59:18	Selamat jalan, maestro. Covid-19 merenggut san...
18	NKRI_NTBSasambo	2021-07-02 23:59:16	SEMUA FASILITAS PEMERINTAH UNTUK RAKYAT DAN PE...
19	falling_don	2021-07-02 23:59:05	Meningkatnya angka covid-19 membuat kebijakan ...

Gambar 4.3 Data Hasil *Crawling*

Pada gambar 4.1 terlihat bahwa hasil dari crawling data terdapat 3 variabel data yang didapat yang terdiri dari variabel *text*, *tweetcreatedts*, dan *username*.

4.2 Pelabelan Data

Pada tahap ini data yang sudah didapat dengan proses crawling data twitter. Data tersebut diseleksi dan diberi pelabelan yaitu kategori positif, negatif, dan netral. Pada tahapan ini penulis juga dibantu ahli dalam memvalidasi data untuk tahapan labeling agar mempermudah dilakukannya labeling data twitter. Data yang diberi label ini digunakan untuk pembuatan sistem. Jadi dari 1000 data yang sudah diambil diseleksi menjadi data yang sudah mewakili menjadi 406 data. Hal ini dilakukan pelabelan dikarenakan membutuhkan waktu yang sangat lama apabila dilakukan pemberian label diseluruh data. Berikut contoh dari pelabelan data yang sudah dilakukan secara manual.

Tabel 4.1 Tabel Pelabelan Data

Kalimat Tweet	Kategori Label
Selamat menikmati liburan ppkm darurat covid 19	Positif
Anggaran Bansos Covid-19 triliunan rp masih minta sumbangan memalukan	Negatif
Jangan ditiru yaa...pejabat selalu kasih contoh jelek, tp terus nyalahin rakyat ga disiplin Prokes Covid-19	Netral

4.3 Pre-procesing Teks

Setelah dilakukan pelabelan data, langkah selanjutnya yaitu preprocesing teks. Didalam tahap ini ada beberapa tahan yaitu proses *cleansing*, *case folding*, penghapusan *stopword* dan *stemming*.

Tahan yang pertama dilakukan adalah *cleansing* yaitu semua karakter *website* atau html dihapus. Hal tersebut dilakukan dikarenakan karkater html tersebut tidak mempunyai makna dalam proses pengklasifikasian nanti. Berikut contoh dari proses *cleansing*.

Tabel 4.2 Tabel Tahapan *Cleansing*

Sebelum <i>Cleansing</i>	Sesudah <i>Cleansing</i>
Disiplin perketat prokes https://t.co/s4nL4zLI2u	Disiplin perketat prokes
Tak Mau Terlantarkan Pasien Covid-19, Anies Ubah JIExpo Jadi Tempat Isolasi https://t.co/0tFAsXOqk0	Tak Mau Terlantarkan Pasien Covid-19, Anies Ubah JIExpo Jadi Tempat
Karena bisa saja datanya disalahgunakan untuk kejahatan. https://t.co/QNrR5ZLul3	Karena bisa saja datanya disalahgunakan untuk kejahatan

Tahapan selanjutnya adalah *case folding*. Pada tahap ini semu huruf dalam kalimat diubah kebentuk standar atau *lower case*. Selain huruf yang diubah karakter selain huruf juga tanda baca dan angka dihapus. Berikut contoh dari tahapan *case folding*

Tabel 4.3 Tabel Tahapan *Case Folding*

Sebelum <i>Case Folding</i>	Sesudah <i>Case Folding</i>
@NajwanHalimi Memang sangat mudah untuk mengatakan yang buruk kepada orang lain.	najwanhalimi memang sangat mudah untuk mengatakan yang buruk kepada orang lain
Buat kamu yang masih harus beraktivitas di luar rumah, jangan lupa Pesan Ibu, ya! Yuk, sama-sama kita cegah penyebaran Covid-19!	buat kamu yang masih harus beraktivitas di luar rumah jangan lupa pesan ibu ya yuk sama sama kita cegah penyebaran covid
Jangan ditiru yaa...pejabat selalu kasih contoh jelek, tapi terus nyalahin rakyat ga disiplin Prokes Covid-19	jangan ditiru yaa pejabat selalu kasih contoh jelek tapi terus nyalahin rakyat ga disiplin prokes covid

Tahapan selanjutnya yaitu penghapusan *stopword*. Pada tahap ini, semua kata yang masuk dalam daftar *stopword* bahasa indonesia dihapus. Penghapusan

kata-kata *stopword* ini dilakukan karena kata yang masuk dalam kategori tersebut tidak digunakan . Pada tabel 4.5 menunjukkan perbedaan dari data sebelum dan sesudah di *stopword*.

Tabel 4 4 Tabel Tahapan *Stopword*

Sebelum dilakukan <i>Stopword</i>	Sesudah dilakukan <i>Stopword</i>
najwanhalimi memang sangat mudah untuk mengatakan yang buruk kepada orang lain	najwanhalimi mudah mengatakan buruk orang lain
buat kamu yang masih harus beraktivitas di luar rumah jangan lupa pesan ibu ya yuk sama sama kita cegah penyebaran covid	buat kamu harus beraktivitas luar rumah jangan lupa pesan ibu kita cegah penyebaran covid
jangan ditiru yaa pejabat selalu kasih contoh jelek tapi terus nyalahin rakyat ga disiplin prokes covid	jangan ditiru pejabat kasih contoh jelek nyalahin rakyat disiplin prokes covid

Tahapan terakhir dari *preprocessing* teks yaitu melakukan tokenisasi dan *stemming*. Tokenisasi dilakukan untuk membagi atau partisi yang berupa data kalimat menjadi kata-kata. Dan tahapan *stemming* yaitu kata-kata yang sudah dipartisi diubah kebentuk dasarnya. Hal ini dilakukan agar kata-kata yang ada pada dokumen twitter bisa diberikan bobot dalam masing-masing kata. Dalam proses ini digunakan *library nltk.tokenize* untuk proses tokenisasi dan *library Sastrawi* digunakan untuk proses stemming.

```
# tokenize tweets
tokenizer = TweetTokenizer(preserve_case=False, strip_handles=True, reduce_len=True)
tweet_tokens = tokenizer.tokenize(tweet)

tweets_clean = []
for word in tweet_tokens:
    if (word not in stopwords_indonesia and # remove stopwords
        word not in emoticons and # remove emoticons
        word not in string.punctuation): # remove punctuation
        #tweets_clean.append(word)
        stem_word = stemmer.stem(word) # stemming word
        tweets_clean.append(stem_word)

return tweets_clean
```

Gambar 4.4 Sourcecode Tokenisasi dan *Stemming*

Pada hasil sourcecode pada gambar 4.2 diatas dapat dilihat di tabel 4.6 dibawah ini dengan beberapa sampel teks tweet yang sudah diproses.

Tabel 4.5 Tabel Tahapan *Tokenisasi* dan *Stemming*

Data Sebelum Tokenisasi dan <i>Stemming</i>	Setelah proses Tokenisasi dan <i>Stemming</i>
najwanhalimi mudah mengatakan buruk orang lain	najwanhalimi – mudah – kata – buruk – orang – lain
buat kamu harus beraktivitas luar rumah jangan lupa pesan ibu kita cegah penyebaran covid	buat – kamu – harus – aktivitas – luar – rumah – jangan – lupa – pesan – ibu – kita – cegah – sebar – covid
jangan ditiru pejabat kasih contoh jelek nyalahin rakyat disiplin prokes covid	jangan – tiru – pejabat – kasih – contoh – jelek – salah – rakyat – disiplin – prokes – covid

4.4 Pembobotan Kata

Proses pembobotan yaitu proses dimana setiap kata yang sudah dilakukan stemming diberikan bobot pada tiap katanya berdasarkan tingkat kemunculan kata pada suatu dokumen. Pada pembobotan kata ini ada dua tahapan perhitungan yaitu perhitungan TF dan TF-IDF. Metode TF termasuk metode yang sering digunakan dengan cara menghitung bobot tiap kata dengan menghitung banyaknya kemunculan kata dalam suatu dokumen tweet.

Tahapan pembobotan diawali dengan perhituangen nilai *TF* (*term frequency*), sebagai contoh berikut beberapa dataset ujicoba yang berisi dari kata-kata yang telah dilakukan stemming. Terdapat 3 dokumen yang akan dilakukan perhitungan, berikut contoh dokumen yang telah dilakukan stemming.

Tabel 4.6 Tabel Data

Dokumen	Teks
D1	covid ajar sakit dekat
D2	covid jakarta tambah sakit
D3	tambah sakit parah

Hal yang harus dilakukan selanjutnya yaitu menyatukan term yang ada di seluruh dokumen menjadi satu sebagai representasi seluruh dokumen. Kemudian dilakukan perhitungan frequency kemunculan term yang ada apada tiap dokumen, misalkan term “covid” pada D1 sebanyak 1 kali maka TF dari term “covid” yaitu 1.

Tabel 4.7 Tabel Term Frequency

	D1	D2	D3
Covid	1	1	0
Ajar	1	0	0
Sakit	1	1	1
Jakarta	0	1	0
Tambah	0	1	1
Parah	0	0	1

Tahapan selanjutnya yaitu nilai Tf yang sudah didapatkan digunakan dalam perhitungan DF (document frequency). DF didapatkan dari jumlah kemunculan term pada dokumen semisal term “covid” yang muncul sebanyak 2 kali, pada D1 dan D2 maka nilai dari DF pada term “covid” yaitu 2.

Tabel 4.8 Tabel DF (Documen Frequency)

	D1	D2	D3	DF
Covid	1	1	0	2

Ajar	1	0	0	1
Sakit	1	1	1	3
Jakarta	0	1	0	1
Tambah	0	1	1	2
Parah	0	0	1	1

Selanjutnya yaitu melakukan perhitungan dengan TF-IDF. Metode pembobotan kata ini merupakan pengembangan dari tahapan perhitungan metode TF dikarenakan setelah dilakukan pembobotan kata TF maka kemudian bobot dari setiap kata tersebut dikalikan dengan IDF (banyak kemunculan kata pada dokumen). Pada metode ini diketahui jika semakin jarang kemunculan sebuah kata pada dokumen, maka hal ini akan membuat nilai IDF nya semakin besar. Sebagai contoh kita gunakan *term* “beli”, dengan jumlah total seluruh dokumen yaitu 3 dengan nilai *DF term* “beli” yaitu 2. Kemudian dilakukan perhitungan dengan menggunakan rumus berikut.

$$idf(t) = \ln \frac{D}{df(t)} + 1$$

$$idf(covid) = \ln \frac{3}{2} + 1 = 1.4054651$$

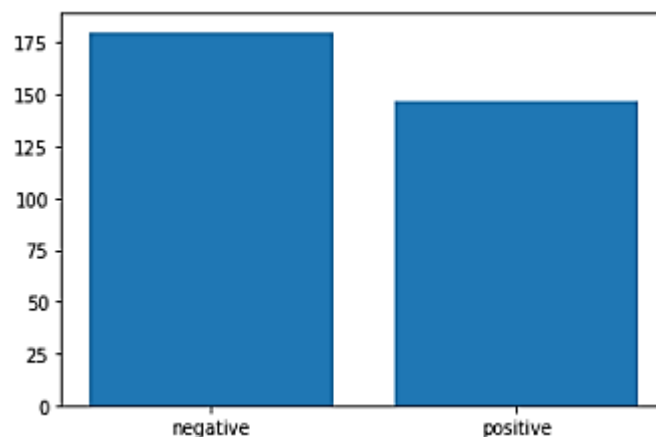
Tabel 4.9 Tabel IDF

	D1	D2	D3	DF	IDF
Covid	1	1	0	2	1.405
Ajar	1	0	0	1	2.098
Sakit	1	1	1	3	1
Jakarta	0	1	0	1	2.098
Tambah	0	1	1	2	1.405
Parah	0	0	1	1	2.098

4.5 Klasifikasi Dengan *Logistic Regression*

Tahapan selanjutnya adalah melakukan klasifikasi data *tweet* dengan yang sudah diberi bobot pada data tweet dari tiap kata menggunakan pembobotan kata TF-IDF. Maka selanjutnya yaitu membuat sistem klasifikasi data tweet menggunakan *Logistic Regression*. Output dari model prediksi yang dibuat dengan *Logistic Regression* ini akan menunjukkan nilai performa akurasi sistem yang telah dibuat.

Hal pertama yang dilakukan adalah membagi dataset twitter menjadi data training dan data testing. Perbandingan jumlah data yang sering digunakan dalam membagi jumlah data yaitu 20% untuk data testing dan 80% untuk data training. Data yang diperoleh pada tahapan *crawling* yaitu berjumlah 1000 data, setelah dilakukan penghapusan duplikat pada data tweet maka jumlah data menjadi 406 data tweet. Data tersebut diberi label positif dan negatif dengan data positif berjumlah 147 data, data berlabel negatif berjumlah 180 data sedangkan data yang tersisa dihapus karena berlabel netral tidak diperlukan dalam menentukan sentimen positif atau negatif. Berikut tampilah data grafik dari label :



Gambar 4.5 Data Grafik Label

Dari tampilan gambar 4.3 bahwa distribusi kelas membantu dalam klasifikasi teks. Dimana dalam data contoh 95% data berada dalam satu kelas dan 5% sisanya dibagi diantara 5 kelas lainnya. Apabila tidak dilakukan pengolahan data, model hanya akan belajar menebak kelas 95% sepanjang waktu dan akan benar 95 % dari waktu pada data yang akan digunakan.

Dari data yang sudah ada dilakukan vectorisasi menjadi variabel *x* dan *y*. Dimana *x* merupakan data teks tweet dan *y* merupakan data label. Berikut tampilan *sourcecode*nya :

```
x = df.Text.values
y = df.Label.values
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.20, random_state=32)
```

Gambar 4.5 Sourcecode dari pemberian variabel X dan Y

Dalam klasifikasi dataset *vector* akan terdiri dari indeks setiap kata yang digunakan dalam dataset pelatihan dan dapat memeriksa tampilannya hanya dengan mencetak tweet pertama baik seperti sebelumnya maupun setelah melakukan vectorisasi. Berikut tampilan *sourcecode* dan hasilnya :

```
vectorizer = CountVectorizer()
vectorizer.fit(x_train)

X_train = vectorizer.transform(x_train)
X_test = vectorizer.transform(x_test)

print(x_train[0], '\n', X_train[0])
```

Gambar 4.6 Sourcecode Vectorisasi Tahap 1

Hasil dari dijalankannya *sourcecode* :

```

tekan sebar covid kabupaten gresik rumah tahan rutan kelas iib banjarsari cerme semprot disinfektan berita news kini jatim covi
d
(0, 166) 1
(0, 209) 1
(0, 301) 1
(0, 328) 2
(0, 386) 1
(0, 524) 1
(0, 604) 1
(0, 692) 1
(0, 747) 1
(0, 800) 1
(0, 843) 1
(0, 1131) 1
(0, 1436) 1
(0, 1439) 1
(0, 1479) 1
(0, 1504) 1
(0, 1625) 1
(0, 1662) 1

```

Gambar 4.7 Hasil Vectorisasi Tahap 1

Pada sebuah matriks terkadang jarang yang dilihatkan tidak sesuai dengan indeks kata dan jumlah dalam data *tweet* dan data yang ditampilkan tersebut tidak dalam urutan yang sama seperti didalam data *tweet* serta dapat memeriksa nilai yang sesuai menggunakan kosakata dari *vectorizer*.

```

import re

d = ",.!?/&-:;@'..."
"["+d.join(d)+"]"

s = x_train[0]
s = ' '.join(w for w in re.split("["+d.join(d)+"]", s) if w)

for i in s.split():
    if len(i)>1: print(i, vectorizer.vocabulary_[i.lower()])

```

Gambar 4.8 Sourcecode Vectorisasi Tahap 2

Hasil dari dijalankannya *sourcecode* :

```

tekan 1662
sebar 1479
covid 328
kabupaten 747
gresik 524
rumah 1436
tahan 1625
rutan 1439
kelas 800
iib 604
banjarsari 166
cerme 301
semprot 1504
disinfektan 386
berita 209
news 1131
kini 843
jatim 692
covid 328

```

Gambar 4.9 Hasil Vectorisasi Tahap 2

Tahapan selanjutnya adalah *modeling* menggunakan klasifikasi *Logistic Regression*. Pada tahapan menggunakan librari *Scikit-Learn* yang mana membantu dalam keperluan perhitungan *data science* dengan klasifikasi *Logistic Regression* ini mendapatkan akurasi sebesar 74%. Berikut *sourcecode* dari klasifikasi *Logistic Regression* beserta *score* akurasi yang didapat dari perujian dengan *Logistic Regression* :

```

1 classifier = LogisticRegression(max_iter=1000)
2 classifier.fit(X_train, y_train)
3
4 score = classifier.score(X_test, y_test)
5
6 print("Accuracy:", score)

```

Accuracy: 0.7424242424242424

Gambar 4.10 Sourcecode Klasifikasi dan Akurasi

4.6 Evaluasi Sistem

Pada atahan ini selanjutnya yaitu melakukan uji evaluasi terhadap sistem klasifikasi yang sudah digunakan. Evaluasi ini berupa pengujian terhadap sistem klasifikasi yang sudah digunakan dalam pengklasifikasian data tweet, sehingga diketahui kualitas dari sistem klasifikasi yang telah dibuat dari data yang sudah ada. Pengujian yang digunakan yaitu menggunakan *Confusion Matrix*. *Confusion Matrix* ini bisa digunakan dalam pengukuran peforma sistem klasifikasi dengan menghitung akurasi, presisi, *recall*, dan *f-measure*. Perhitungan – perhitungan ini sangat berguna untuk mengetahui peforma dari sistem klasifikasi yang dibuat. Oleh karena itu *confusion matrix* merupakan sebuah matriks yang digunakan untuk membandingkan hasil klasifikasi dengan data yang sebenarnya (asli).

Tabel 4.10 *Confusion Matrix*

Prediksi Aktual	Positif	Negatif
Positif	31	6
Negatif	11	18

Dari tabel diatas didapatkan hasil dari *True Positive*, *False Positive*, *True Negative*, dan *False Negative* :

$$TP (True Positive) = 31$$

$$FP (False Positive) = 11$$

$$TN (True Negative) = 18$$

$$FN (False Negative) = 6$$

Berikut ini merupakan hasil dari perhitungan pada akurasi, *presisi*, *recall*, dan *f-measure* berdasarkan tabel 4.7 :

$$\begin{aligned} Akurasi &= \frac{TP + TN}{TP + FP + FN + TN} \cdot 100\% \\ &= \frac{31 + 18}{31 + 11 + 6 + 18} \cdot 100\% \\ &= 0,74242 \cdot 100\% \\ &= 74,2\% \end{aligned}$$

$$\begin{aligned} Presisi\ Positif &= \frac{TP}{TP + FP} \cdot 100\% \\ &= \frac{31}{31 + 11} \cdot 100\% \\ &= 0,7380 \cdot 100\% \\ &= 73,8\% \end{aligned}$$

$$\begin{aligned} Presisi\ Negatif &= \frac{TN}{TN + FN} \cdot 100\% \\ &= \frac{18}{18 + 6} \cdot 100\% \\ &= 0,75 \cdot 100\% \\ &= 75\% \end{aligned}$$

$$Rata - Rata\ Presisi = \frac{Presisi\ Positif + Presisi\ Negatif}{2}$$

$$= 74,5 \%$$

$$\begin{aligned} \text{Recall Positif} &= \frac{TP}{TP + FN} \cdot 100\% \\ &= \frac{31}{31 + 6} \cdot 100\% \\ &= 0,8378 \cdot 100\% \\ &= 83,7\% \end{aligned}$$

$$\begin{aligned} \text{Recall Negatif} &= \frac{TN}{TN + FP} \cdot 100\% \\ &= \frac{18}{18 + 11} \cdot 100\% \\ &= 0,6206 \cdot 100\% \\ &= 62\% \end{aligned}$$

$$\begin{aligned} \text{Rata - Rata Recall} &= \frac{\text{Recall Positif} + \text{Recall Negatif}}{2} \\ &= 73\% \end{aligned}$$

$$\begin{aligned} F - \text{Measure} &= 2 \cdot \left(\frac{\text{Presisi} \cdot \text{Recall}}{\text{Presisi} + \text{Recall}} \right) \\ &= 2 \cdot \left(\frac{74,5\% \cdot 73\%}{74,5\% + 73\%} \right) \\ &= 73\% \end{aligned}$$

4.7 Visualisasi Data

Setelah pengolahan dan klasifikasi data *tweet*, tahapan selanjutnya pada penelitian ini adalah memvisualisasikan data tersebut agar lebih mudah dipahami. Langkah ini diambil dengan membuat *wordcloud* yang berfungsi untuk mengetahui kata-kata mana yang memiliki kemunculan lebih sering sehingga dapat berpengaruh dalam memahami sentimen dan pembuatan model klasifikasi. *Wordcloud* merupakan tampilan yang menunjukkan frekuensi kata dalam sebuah dokumen dan pada tahap ini akan dilakukan penggunaan *wordcloud* dari dokumen data *tweet* positif dan negatif.

Dari gambar 4.11 dapat dilihat bahwa kata-kata yang memiliki frekuensi kemunculan pada data berlabel negatif yaitu covid, perintah, orang, tangan, tinggal, dan lain-lain. Selain itu juga terdapat kata-kata yang sering muncul selama masa pandemi ditahun ini seperti pphk, darurat, rakyat, infeksi, dan lain sebagainya. Dengan melihat *wordcloud* diatas dapat diketahui bahwa sentimen ataupun tweet yang sering muncul dalam data yang berlabel negatif menunjukkan kecemasan maupun kesedihan atas situasi pandemi di Indonesia.



Gambar 4.12 Wordcloud Berlabel Positif

Pada gambar 4.12 yang merupakan visualisasi dari *wordcloud* yang menunjukkan kata-kata yang sering muncul pada data tweet berlabel positif. Pada gambar 4.12 juga terlihat bahwa kata-kata dengan kemunculan paling tinggi yaitu covid, indonesia, vaksin, sebar, sehat, ppkm, dan lain sebagainya. Dari kemunculan kata-kata tersebut menunjukkan harapan dan keinginan dari masyarakat Indonesia khususnya untuk melawan covid ini sehingga dapat hidup nyaman dan pastinya sehat.

BAB 5. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan penelitian yang sudah dilakukan ini, Analisis Sentimen Pada Penggunaan *Hashtag* Covid -19 di Media Sosial *Twitter* dengan menggunakan klasifikasi Logistic Regression dapat ditarik kesimpulan sebagai berikut :

1. Pada tahapan preproses teks yaitu stemming menggunakan *library sastrawi* masih banyaknya kata yang tidak bisa dikembalikan ke kata dasar dikarenakan pengetikan yang tidak baku atau typo, sehingga pada data yang diambil yang sudah mengalami *text processing* dan bisa berakibat pembobotan kata yang kurang maksimal
2. Dari proses pengambilan data, data yang didapat berjumlah 1000 data, namun setelah dilakukan penghapusan data duplikat, dataset yang tersisa berjumlah 406 data dengan data netral di hapus dan menjadi 180 data berlabel negatif dan 147 data berlabel positif.
3. Model atau Metode Klasifikasi *Logistic Regression* merupakan metode klasifikasi yang digunakan dalam penelitian ini dan metode tersebut digunakan mengklasifikasi sentimen dari hashtag “COVID-19”. Tujuan dari penelitian ini untuk mengetahui kecenderungan sentimen atau opini masyarakat terhadap topik covid-19 di media sosial twitter di Indonesia. Pada penggunaan klasifikasi *Logistic Regression* di dapat akurasi 74 % dan hasil dari perhitungan *presisi*, *recall*, dan *f-measure* yaitu 74,5%, 73%, dan 73%

5.2 Saran

Pada penelitian tugas akhir ini, terdapat beberapa saran yang bisa digunakan pertimbangan untuk penelitian selanjutnya :

1. Perlu dilakukan pengambilan data yang lebih banyak agar pada saat tahapan preproses dan pembobotan lebih maksimal.

2. Pada saat tahapan stemming perlu melengkapi kata – kata lagi dengan dibantu librari sastrawi. Sehingga sewaktu melakukan proses *stemming* , kata-kata yang tidak baku otomatis berubah menjadi kata baku secara otomatis dan lebih baik lagi.
3. Untuk tahapan klasifikasi *Logistic Regression* dengan vektorisasi dengan hasil akurasi 74% bisa dilakukan pendekatan yang lain yang performanya lebih baik.

DAFTAR PUSTAKA

- (Rizaldi & Faraby, 2020)Ahp, P. T. (2019). *Analisis Kepribadian Melalui Twitter Menggunakan Metode Logistic Regression dengan*. 6(2), 9667–9682.
- Aldean, M. Y., Hilmawan, M. D., Indriyati, R., & Lasama, J. (2019). Analisa Relevansi Tweet terhadap Hashtag dengan Metode Logistic Regression. *Centive*, 2(1), 32–38.
- Ardhiansyah, M. N., Umar, R., & Sunardi. (2019). Analisis Sentimen pada Twitter Menggunakan Metode Support Vector Machine. *Seminar Nasional Teknologi Fakultas Teknik Universitas Krisnadwipayana*, 1(1), 739–742.
- Cahyono, Y. (2017). Analisis Sentiment pada Sosial Media Twitter Menggunakan Naïve Bayes Classifier dengan Feature Selection Particle Swarm Optimization dan Term Frequency. *Jurnal Informatika Universitas Pamulang*, 2(1), 14. <https://doi.org/10.32493/informatika.v2i1.1500>
- Dwi, E., Sari, N., Statistika, D., Matematika, F., & Data, S. (2019). Analisis Sentimen Nasabah Pada Layanan Perbankan Menggunakan Metode Regresi Logistik Biner , Naïve Bayes Classifier (NBC), dan Support Vector Machine (SVM). *Jurnal Sains Dan Seni Its*, 8(2), 177.
- Fauzi, A., Akbar, M. F., & Asmawan, Y. F. A. (2019). Sentimen Analisis Berinternet Pada Media Sosial dengan Menggunakan Algoritma Bayes. *Jurnal Informatika*, 6(1), 77–83. <https://doi.org/10.31311/ji.v6i1.5437>
- Fauziyyah, A. K. (2020). Analisis Sentimen Pandemi Covid19 Pada Streaming Twitter Dengan Text Mining Python. *Jurnal Ilmiah SINUS*, 18(2), 31. <https://doi.org/10.30646/sinus.v18i2.491>
- Imamah, & Rachman, F. H. (2020). Twitter sentiment analysis of Covid-19 using term weighting TF-IDF and logistic regresion. *Proceeding - 6th Information Technology International Seminar, ITIS 2020*, 238–242. <https://doi.org/10.1109/ITIS50118.2020.9320958>
- Islam, U., Sunan, N., Yogyakarta, K., Sarjana, G., Satu, S., & Komunikasi, I. (2016). *Pengaruh Penggunaan Media Sosial Thp Penyebaran Infokes*.
- Junco, R., Elavsky, C. M., & Heiberger, G. (2013). Putting twitter to the test: Assessing outcomes for student collaboration, engagement and success. *British Journal of Educational Technology*, 44(2), 273–287. <https://doi.org/10.1111/j.1467-8535.2012.01284.x>
- Lesmana, P. I. (2013). Analisis sentimen ..., Pekik Indra Lesmana, Fasilkom UI, 2013. *Karya Akhir, Universitas Indonesia*.
- Murnawan, M. (2017). Pemanfaatan Analisis Sentimen Untuk Pemingkatan Popularitas Tujuan Wisata. *Jurnal Penelitian Pos Dan Informatika*, 7(2), 109. <https://doi.org/10.17933/jppi.2017.070203>
- Nurjanah, W. E., Perdana, R. S., & Fauzi, M. A. (2017). Analisis Sentimen

Terhadap Tayangan Televisi Berdasarkan Opini Masyarakat pada Media Sosial Twitter menggunakan Metode K-Nearest Neighbor dan Pembobotan Jumlah Retweet. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya*, 1(12), 1750–1757.
<https://doi.org/10.1074/jbc.M209498200>

Nuryanto, A. (2021). *COVID-19 MENGGUNAKAN METODE LOGISTIC REGRESSION*. 5(2), 234–241.

Ratnasari, V. (2017). *Pengoptimalan Naïve Bayes Dan Regresi Logistik Menggunakan Algoritma Genetika Untuk Data Klasifikasi*. 86.

Rizaldi, M. N., & Faraby, S. Al. (2020). *Klasifikasi Argument Pada Teks dengan Menggunakan Metode Multinomial Logistic Regression Terhadap Kasus Pemindahan Ibu Kota Indonesia di Twitter*. 4(L), 904–913.
<https://doi.org/10.30865/mib.v4i4.2348>

Salim, A. (2019). *Optimalisasi Regresi Logistik Pada Proses Klasifikasi Menggunakan Algoritma Genetika*. 6(2), 50–55.
<https://doi.org/10.25047/jtit.v6i2.109>

Sari, F. V., & Wibowo, A. (2019). Analisis Sentimen Pelanggan Toko Online Jd. Id Menggunakan Metode Naïve Bayes Classifier Berbasis Konversi Ikon Emosi. *Simetris: Jurnal Teknik Mesin, Elektro Dan Ilmu Komputer*, 2(2), 681–686.

Zenico, R., Setiawan, E. B., & Nugraha, F. N. (2019). Prediksi Big Five Personality dengan Term Frequency Inverse Document Frequency (TF – IDF) Menggunakan Metode Logistic Regression pada Pengguna Twitter. -- *Proceeding of Engineering*, 6(2), 9939–9945.

Nirwana. S.R.A. 2015. Regresi Logistik Multinomial dan Penerapannya dalam Menentukan Faktor yang Berpengaruh pada Pemilihan Program Studi di Jurusan Matematika UNM. *Skripsi*. Universitas Negeri Makassar. Makassar.

LAMPIRAN

1. Validasi Data



KEMENTERIAN PENDIDIKAN KEBUDAYAAN RISET DAN
TEKNOLOGI

POLITEKNIK NEGERI JEMBER
JURUSAN TEKNOLOGI INFORMASI

Jl. Mastrip PO.BOX 164 Telp : 333532-333534 Fax 333531 Jember 68101

SURAT PERNYATAAN

Yang bertanda tangan di bawah ini :

Nama : Fanny Martha Apriyanti, S.Pd

Jabatan : Guru Kelas di SDN Ampel 03 Wuluhan

Menyatakan dengan sebenar-benarnya bahwa mahasiswa dibawah ini :

Nama : Yuniar Fabi Putra

NIM : E41171845

Judul Skripsi : Analisis Sentimen Pada Penggunaan *Hashtag* COVID-19 di Media Sosial *Twitter*

Jurusan : Teknologi Informasi

Progam Studi : Teknik Informatika Kampus Bondowoso

Asal Studi : Politeknik Negeri Jember

Sehubungan dengan skripsi yang berjudul "**Analisis Sentimen Pada Penggunaan *Hashtag* COVID-19 di Media Sosial *Twitter***". Menyatakan bahwa :

Data-data meliputi tweet tentang *hashtag* COVID-19 pada *Twitter* yang telah dilakukan perbaikan ejaan kata dan telah dilabelkan benar-benar valid tanpa ada perubahan yang mempengaruhi makna asli dari data yang didapatkan.

Demikian pernyataan ini dibuat dengan sesungguhnya dan sebenar-benarnya.

Jember, 29 Juli 2021

Mengetahui,

Fanny Martha Apriyanti, S.Pd

2. Email dari pihak Twitter Developer

