

# **ANALISIS SENTIMEN VAKSIN COVID-19 BERDASARKAN LOKASI DI INDONESIA MENGGUNAKAN METODE *LEXICON* BASED DAN *CNN***

## **LAPORAN TUGAS AKHIR**

Diajukan untuk memenuhi kelulusan matakuliah Tugas Akhir  
Pada Program Studi D4 Teknik Informatika



**Dibuat Oleh :**

**1.17.4.021     Muhammad Fahmi**

**PROGRAM DIPLOMA IV TEKNIK INFORMATIKA**

**POLITEKNIK POS INDONESIA**

**BANDUNG**

**2021**

***COVID-19 VACCINE SENTIMENT ANALYSIS BASED  
ON LOCATION IN INDONESIA USING  
LEXICON BASED AND CNN METHOD***

***FINAL REPORT***

*Submitted to fulfil the graduation of the Final Report  
In Applied Bachelor Program Of Informatic Engineering*



*Created by,*

***1.17.4.021     Muhammad Fahmi***

***APPLIED BACHELOR PROGRAM OF INFORMATICS ENGINEERING  
POLITEKNIK POS INDONESIA***

***BANDUNG***

***2021***

## ABSTRAK

Wabah penyakit virus Covid-19 telah ditetapkan sebagai pandemi *global* oleh *WHO* pada tanggal 11 maret 2020 lalu dalam situs resminya. Melihat pesatnya penyebaran Covid-19 dan bahaya yang akan muncul jika tidak segera di tangani, maka salah satu cara yang dilakukan untuk mencegah penyebaran virus ini adalah dengan mengembangkan vaksin. Tetapi, banyak respon dan opini yang diberikan masyarakat diberbagai media sosial. Diantaranya media sosial *Twitter*. Oleh sebab itu, penulis akan melakukan analisis sentimen positif, negatif serta netral pengguna *twitter* tentang vaksin Covid-19 pada media sosial *Twitter* berdasarkan persebaran provinsi di Indonesia. Penelitian ini menggunakan bahasa Python dan Lexicon Based dan *Convolutional Neural Network* (CNN) dari *Deep Learning*. Dengan demikian, kesimpulan dapat ditarik sebuah kesimpulan dalam bentuk visualisasi *dashboard*. Penelitian ini akan membandingkan beberapa *word embedding*, diantaranya *FastText*, *Word2vec* dan *GloVe*. Akurasi pada *Lexicon Based* sebesar 95% dan untuk *CNN+FastText* sebesar 97%.

**Kata Kunci : Vaksin, Sentimen, *Lexicon Based*, CNN, FastText**

## ***ABSTRACT***

*On its official website, the outbreak of the Covid-19 virus disease was declared a global pandemic by WHO on March 11, 2020. Seeing the rapid spread of Covid-19 and the dangers that will arise if it is not handled immediately, one way to prevent the spread of this virus is to develop a vaccine. However, many responses and opinions were given by the public on various social media. One of them is Twitter social media. Therefore, the author will analyze Twitter users' positive, negative, and neutral sentiments about the Covid-19 vaccine on Twitter social media based on the distribution of provinces in Indonesia. This study uses Python and Lexicon Based and Convolutional Neural Network (CNN) from Deep Learning. Thus, a conclusion can be drawn in the form of dashboard visualization. This study will compare several word embedding, including FastText, Word2vec, and GloVe. The accuracy for Lexicon Based is 95%, and for CNN+FastText, it is 97%.*

***Keywords: Vaccine, Sentiment, Lexicon Based, CNN, FastText***

## **KATA PENGANTAR**

Assalamu'alaikum Warahmatullahi Wabarakatuh.

Puji syukur penulis panjatkan ke hadirat Allah SWT karena hanya dengan rahmat dan hidayahnya, laporan tugas akhir ini dapat terselesaikan tanpa halangan berarti. Keberhasilan dalam menyusun laporan tugas akhir ini pasti ada beberapa bantuan dari berbagai pihak untuk membantu laporan tugas akhir ini. Oleh karena itu, dalam kata pengantar ini, penulis sangat banyak mengucapkan terima kasih kepada:

1. M. Yusril Helmi Setyawan, S.Kom., M.Kom. Selaku Ketua Program Studi DIV Teknik Informatika Politeknik Pos Indonesia,
2. Mohamad Nurkamal Fauzan, S.T., M.T. Selaku Koordinator tugas akhir tahun 2020/2021,
3. Rolly Maulana Awangga, S.T., M.T. dan Roni Andarsyah, S.T., M.Kom. Selaku dosen pembimbing penulis pada tugas akhir tahun 2020/2021,
4. Orang Tua, yang selalu memberi dukungan dan doa,
5. Teman - teman yang juga memberi dukungan dan doa.

Penulis menyadari bahwa penyusunan laporan tugas akhir ini masih banyak kesalahan yang perlu untuk di tingkatkan kepannya. Kritik dan saran dapat ditujukan kepada penulis. Dengan diselesaikannya laporan tugas akhir, penulis berharap semoga laporan ini dapat berjalan dengan lancar saat proses pengerjaannya nanti.

Wassalamu'alaikum Warahmatullahi Wabarakatuh.

Bandung, 18 Agustus 2021

Penulis

# DAFTAR ISI

ABSTRAK .....	i
<i>ABSTRACT</i> .....	ii
KATA PENGANTAR.....	iii
DAFTAR ISI .....	iv
DAFTAR GAMBAR .....	vi
DAFTAR TABEL .....	vii
DAFTAR SIMBOL.....	viii
DAFTAR SINGKATAN.....	ix
DAFTAR LAMPIRAN .....	x
BAB I PENDAHULUAN .....	I-1
1.1. Latar Belakang .....	I-1
1.2. Identifikasi Masalah .....	I-3
1.3. Tujuan dan Manfaat .....	I-3
1.4. Ruang Lingkup.....	I-3
BAB II LANDASAN TEORI .....	II-4
2.1. Teori .....	II-4
2.1.1. <i>Covid-19</i> .....	II-4
2.1.2. Vaksin .....	II-5
2.1.3. <i>Twitter</i> .....	II-5
2.1.4. Analisis Sentimen .....	II-6
2.1.5. Metode <i>Lexicon Based</i> .....	II-7
2.1.7. <i>Python</i> .....	II-10
2.2. Tinjauan Pustaka .....	II-11
BAB III14 OBJEK STUDI .....	III-14
3.1. Material .....	III-14
3.2. Teknologi .....	III-14
BAB IV METODOLOGI PENELITIAN .....	IV-19
4.1. Diagram Alur Metodologi Penelitian.....	IV-19
4.2.1. <i>Business Understanding</i> .....	IV-20
4.2.2. <i>Data Understanding</i> .....	IV-20
4.2.3. <i>Data Preparation</i> .....	IV-20

4.2.4. <i>Modelling</i> .....	IV-20
4.2.4.1. Scraping Data .....	IV-21
4.2.4.2. Preprocessing/Cleaning Data .....	IV-22
4.2.4.3. Sentiment Analysis .....	IV-23
4.2.4.4. Output.....	IV-27
4.2.5. <i>Evaluation</i> .....	IV-27
BAB V IMPLEMENTASI DAN PENGUJIAN .....	V-28
5.1. Lingkungan Implementasi.....	V-28
BAB VI KESIMPULAN DAN SARAN .....	VI-46
DAFTAR PUSTAKA .....	47
LAMPIRAN .....	11

## DAFTAR GAMBAR

Gambar 2.1. Contoh <i>Tweet</i> Pada <i>Twitter</i> .....	II-6
Gambar 2.2. Contoh Klasifikasi <i>Lexicon Based</i> .....	II-7
Gambar 2.3. Arsitektur <i>CNN</i> .....	II-9
Gambar 2.4. Most Popular Programming Language.....	II-11
Gambar 4.1. Metodologi Penelitian .....	IV-19
Gambar 4.2. Diagram Alir Penelitian.....	IV-21
Gambar 4.3. Diagram Alir Model Proses Pengumpulan Data <i>Twitter</i> .....	IV-22
Gambar 4.4. Arsitektur <i>CNN</i> untuk Analisis Sentimen.....	IV-22
Gambar 5.1. <i>Summary GPU Google Colab</i> .....	V-29
Gambar 5.2. <i>Labelling</i> Data Manual.....	V-30
Gambar 5.3. Langkah-Langkah <i>Pre-processing Data</i> .....	V-31
Gambar 5.4. <i>Pie Chart Labelling</i> Manual.....	V-33
Gambar 5.5. Contoh Hasil <i>Lexicon Based</i> .....	V-34
Gambar 5.6. Contoh Hasil <i>Lexicon Based</i> Setelah di sentimen.....	V-34
Gambar 5.7. <i>Classification Report Lexicon Based</i> .....	V-35
Gambar 5.8. Confussion Matrix <i>Lexicon Based</i> .....	V-35
Gambar 5.9. <i>Pie Chart Hasil Lexicon Based</i> .....	V-36
Gambar 5.10. Contoh Arsitektur <i>CNN</i> .....	V-38
Gambar 5.11. Contoh Word Embedding Fasttext .....	V-38
Gambar 5.10. Contoh Arsitektur <i>CNN</i> .....	V-38
Gambar 5.11. Contoh Word Embedding Fasttext .....	V-38



## DAFTAR TABEL

Tabel 1. Perbandingan Penelitian Sebelumnya Dengan Penelitian Yang Akan Dilakukan..... II-

**Error! Bookmark not defined.**

Tabel 1. Perbandingan Penelitian Sebelumnya Dengan Penelitian Yang Akan Dilakukan..... 11

Tabel 1. Perbandingan Penelitian Sebelumnya Dengan Penelitian Yang Akan Dilakukan..... 11

Tabel 1. Perbandingan Penelitian Sebelumnya Dengan Penelitian Yang Akan Dilakukan..... 11

Tabel 1. Perbandingan Penelitian Sebelumnya Dengan Penelitian Yang Akan Dilakukan..... 11

Tabel 1. Perbandingan Penelitian Sebelumnya Dengan Penelitian Yang Akan Dilakukan..... 11

Tabel 1. Perbandingan Penelitian Sebelumnya Dengan Penelitian Yang Akan Dilakukan..... 11

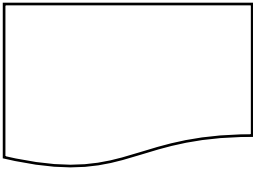
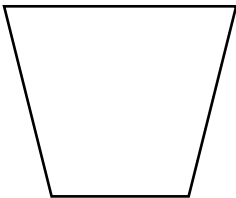

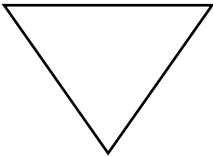
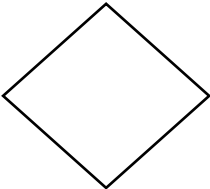
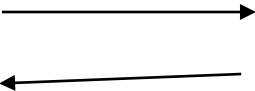

Tabel 1. Perbandingan Penelitian Sebelumnya Dengan Penelitian Yang Akan Dilakukan..... 11

Tabel 1. Perbandingan Penelitian Sebelumnya Dengan Penelitian Yang Akan Dilakukan..... 11

Tabel 1. Perbandingan Penelitian Sebelumnya Dengan Penelitian Yang Akan Dilakukan..... 11

Tabel 1. Perbandingan Penelitian Sebelumnya Dengan Penelitian Yang Akan Dilakukan..... 11

## DAFTAR SIMBOL

NO	SIMBOL	NAMA	KETERANGAN
1		Dokumen	I.O dalam format yang dicetak
2		Manual Operation	Proses yang terjadi di dalam flowmap
3		Proses Komputerisasi	Merepresentasikan Input data atau Ouput data yang di proses atau informasi
4		Arsip Manual	Penyimpanan yang dapat diakses oleh komputer secara langsung
5		Decision	Menunjukkan pilihan keputusan
6		Aliran	Menunjukkan data-data yang mengalir pada sistem
7		Direct Data	Merupakan penyimpanan database

## DAFTAR SINGKATAN

1	<i>NLP</i>	: <i>Natural Language Processing</i>
2	<i>API</i>	: <i>Application Programming Interface</i>
3	<i>HTML</i>	: <i>Hypertext</i>
4	<i>TF-IDF</i>	: <i>Term Frequency-Inverse Document Frequency</i>
5	<i>CSV</i>	: <i>Comma-Separated Values</i>
6	<i>CRISP-DM</i>	: <i>Cross-Industry Standard Process for Data Mining</i>
7	<i>CNN</i>	: <i>Convolutional Neural Network</i>

## **DAFTAR LAMPIRAN**

Lampiran 1. Bukti Kartu Bimbingan (KAMBING).....	50
Lampiran 2. Bukti Pengecekan Plagiarisme Online.....	52
Lampiran 3. Surat Pernyataan .....	52
Lampiran 4. Bukti Sumbit Jurnal .....	54

# BAB I

## PENDAHULUAN

### 1.1. Latar Belakang

Wabah penyakit *covid-19* telah ditetapkan sebagai pandemi *global* oleh *World Health Organization* atau yang biasa disingkat dengan *WHO*, ditetapkan pada tanggal 11 Maret 2020 lalu didalam situs resminya (Ward & del Rio, 2020). Ternyata sangat banyak aspek kehidupan yang menjadi berubah dikarenakan virus *covid-19* atau yang biasa disebut dengan virus corona (Pragholapati, 2020). Khususnya di Indonesia, Presiden Joko Widodo sudah mengumumkan kasus pertama terjadi pada 2 Maret 2020 lalu kasus tersebut mengenai 2 orang WNI asal Depok, Jawa barat (Chadijah et al., 2020). Berdasarkan informasi yang penulis dapatkan sampai tanggal 6 Maret 2021 total sudah ada 1.373.836 kasus di Indonesia dengan menekan angka kematian sampai 37.154 jiwa. Penulis melakukan riset tentang penyebaran *Covid-19* yang sangat pesat dan sangat bahaya jika tidak segera ditangani, salah satu langkah yang diambil oleh pemerintah yaitu dengan mengembangkan vaksin, hal itu dituangkan didalam kutipan (Xu et al., 2020). Vaksin tidak hanya melindungi seseorang yang telah divaksinasi dari virus *covid-19*, tetapi juga mencegah masyarakat umum dengan mencegah penyebaran penyakit ke seluruh populasi yang ada (Sallam, 2021). Sampai dengan detik ini sudah ditemukan vaksin yang layak uji dan aman untuk di konsumsi masyarakat. Vaksin tersebut diharapkan dapat menghentikan penyebaran penyakit dan mencegahnya di masa yang akan datang. Menanggapi hal tersebut, Pemerintah Indonesia sudah terlibat aktif dalam melakukan vaksinasi, hal tersebut dapat di lihat dari pidato singkat Presiden Joko Widodo pada 5 Oktober 2020 yang secara resmi tertuang dalam Peraturan Presiden atau Perpres Republik Indonesia Nomor 99 Tahun 2020 tentang pengadaan vaksin dan pelaksanaan vaksinasi dalam rangka Penanggulangan Pandemi *Corona Virus Disease* 2019 (*COVID-19*) untuk mengatur kewenangan pemerintah, kementerian/Lembaga dan pada pejabatnya dalam rencana kegiatan vaksinasi (Rachman & Pramana, 2020).

Berbicara vaksinasi sangat banyak pertimbangan dan resiko dari berbagai faktor, seperti adanya respon dan opini publik dalam menanggapi vaksin ini. Masyarakat khususnya para pengguna media sosial *Twitter* telah mendapatkan perhatian khusus karena pengguna *Twitter* dapat dengan mudah menyiarkan informasi tentang pendapat mereka terkait vaksinasi melalui pesan public yang disebut *Tweet* (D'Andrea et al., 2019). Selain informasi atau *Tweet* yang dilakukan oleh pengguna *Twitter*. *Twitter* juga menyimpan lokasi pengguna dan mengetahui *Tweet*

tersebut berasal dari daerah atau kota mana yang ada di Indonesia (Henry, 2021). Seiring berjalannya waktu, data *Twitter* sudah dijadikan sebagai bahan penelitian dari berbagai bidang yang ada, terutama dalam menganalisis sentimen terkait topik tertentu berdasarkan kata kunci pada *Twitter* (Cotfas et al., 2021). Topik vaksin telah menjadi salah satu topik yang dapat menimbulkan sejumlah pertanyaan di media sosial, Sebagian besar pro dan kontra terkait vaksin. Oleh karena itu, sejumlah penelitian telah menganalisis pro dan kontra pada keraguan vaksinasi atau kepercayaan masyarakat pengguna *Twitter* tentang adanya Vaksin di Indonesia (Meena, 2020). Dengan seperti itu penulis akan melakukan penelitian terkait analisis sentimen pengguna *twitter* terhadap vaksin. Penelitian ini mengacu dengan yang dilakukan oleh Fajar Fathur Rachman, Setia Pramana terkait analisis respon masyarakat terhadap wacana vaksinasi dengan cara mengklasifikasikan respon tersebut ke dalam respon positif atau negatif dengan menggunakan metode *Latent Dirichlet Allocation (LDA)*. Hasil dari analisis tersebut menunjukkan bahwa masyarakat lebih banyak memberikan respon positif terhadap wacara vaksin tersebut (30%) dan respon negative (26%) (Rachman & Pramana, 2020).

Berdasarkan Penelitian yang dilakukan oleh Fajar Fathur Rachman, Setia Prama pada Desember 2020 lalu, penulis akan mengembangkan penelitian tersebut menjadi analisis sentimen pengguna *twitter* terhadap vaksin *covid-19* dengan mengklasifikan respon positif, negatif dan netral berdasarkan daerah yang ada di Indonesia dengan menggunakan Metode *Lexicon Based* serta CNN untuk melatih data yang sudah di label secara manual dan otomatis menggunakan *lexicon based* (Taboasda et al., 2021). Metode *Lexicon Based* ialah suatu cara klasifikasi sentimen dengan membuat kamus atau opini terlebih dahulu. Kata-kata yang ada pada kamus tersebut akan digunakan untuk proses identifikasi apakah suatu kalimat mengandung opini atau tidak (Wunderlich & Memmert, 2020). *Convolutional Neural Network (CNN)* adalah salah satu jenis neural network yang biasa digunakan untuk proses pembelajaran mesin. CNN dapat digunakan untuk mendeteksi dan mengenali objek. Seluruh proses pengerjaan dalam penelitian ini menggunakan Bahasa Pemrograman *Python* (Suryadi, 2021). Dengan seperti itu penelitian yang penulis lakukan menggunakan semi *supervised learning* yaitu dengan menggabungkan metode *lexicon based* dengan CNN. Dimana *lexicon based* berguna untuk melakukan labelling otomatis terhadap 5000 data untuk menjadi sebuah dataset training. Dataset training tersebut kemudian akan melatih model yang dilakukan oleh metode CNN. Lalu, model tersebut akan melakukan testing terhadap data sisanya yaitu sekitar 28000 data lagi. Penggunaan Algoritma *Convolutional Neural Network (CNN)* dipilih karena dinilai

lebih baik dalam menangani masalah *Big Data* dan *Machine Learning* sehingga bisa dikembangkan untuk membangun model untuk analisis sentimen.

Hasil yang diharapkan dalam penelitian ini adalah bagaimana bisa mengetahui persebaran data berdasarkan provinsi yang beropini positif, negative atau netral dalam menanggapi vaksin *covid-19*.

## **1.2. Identifikasi Masalah**

Identifikasi masalah pada penelitian ini terkait bagaimana melakukan analisis sentimen untuk mengetahui Pro dan Kontra dari data yang ada di media sosial *Twitter*. Dengan melakukan analisis sentimen maka dapat diketahui daerah atau lokasi mana saja yang positif, negatif dan netral terhadap Program Vaksinasi *Covid-19* yang sedang dilakukan Pemerintah Indonesia.

## **1.3. Tujuan dan Manfaat**

Adapun tujuan dan manfaat pada laporan tugas akhir ini adalah :

- a. Menerapkan analisis sentimen.
- b. Melakukan *scraping* data *Twitter* serta melakukan *cleaning*.
- c. Menerapkan metode *lexicon based* dan *CNN* untuk mengidentifikasi data *Twitter* yang ada.

## **1.4. Ruang Lingkup**

Berdasarkan identifikasi masalah serta tujuan dan manfaat diatas, maka ruang lingkup atau batasan dalam proses penelitian laporan Internship II ini adalah :

- a. Metode yang dipakai untuk analisis sentimen ialah *Lexicon Based* dan *CNN*.
- b. Data yang digunakan hanya yang di *scrap* melalui media sosial *Twitter*.
- c. Hasil analisis hanya berdasarkan daerah yang ada di Indonesia.

## **BAB II**

### **LANDASAN TEORI**

#### **2.1. Teori**

##### **2.1.1. Covid-19**

*Covid-19* adalah suatu virus yang pertama kali ditemukan di china. *COVID-19* pertama kali dilaporkan di Indonesia pada 2 Maret 2020 dalam jumlah dua kasus. Pada 31 Maret 2020 menunjukkan sejumlah kasus yang dikonfirmasi 1.528 kasus dan 136 kasus kematian. 10 Angka kematian *COVID-19* di Indonesia adalah 8,9%, angka tersebut adalah tertinggi di Asia Tenggara (Sagala, 2020). Virus corona ini merupakan virus *RNA* dengan partikel ukuran 120-160 nm (Sitepu & Syafril, 2020). Virus ini terutama menginfeksi hewan, termasuk kelelawar dan unta. sebelum Terjadinya wabah *COVID-19* ada 6 jenis virus corona Dapat menginfeksi manusia yaitu *alphacoronavirus* 229E, *alphacoronavirus* NL63, *betacoronavirus* OC43, *betacoronavirus* HKU1, Penyakit Pernafasan Akut Parah Coronavirus (*SARS-CoV*), dan Pernafasan Timur Tengah Sindrom virus korona (*MERS-CoV*) (Al-Sharif et al., 2021). Corona virus yang merupakan etiologi *COVID-19* termasuk dalam genus *betacoronavirus*. Hasil analisis filogenetik menunjukkan bahwa virus ini masuk dalam subgenus yang sama dengan virus corona menyebabkan wabah Penyakit Pernafasan Akut Parah (*SARS*) 2002-2004 yaitu *Sarbecovirus*.

Urutan *SARS-CoV-2* memiliki kemiripan dengan virus korona diisolasi pada kelelawar, sehingga muncul hipotesis *SARS-CoV-2* berasal dari kelelawar saat itu bermutasi dan menginfeksi manusia. Mamalia dan burung dianggap sebagai reservoir perantara. Dalam kasus *COVID-19*, trenggiling diduga reservoir perantara. Jenis virus corona pada trenggiling adalah genom yang mirip dengan kelelawar coronavirus (90,5%) dan *SARS-CoV-2* (91%). Genom *SARS-CoV-2* sendiri memiliki 89% homologi terhadap virus korona kelelawar ZXC21 dan 82% melawan *SARS-CoV*. Hasil pemodelan komputer menunjukkan *SARS-CoV-2* memiliki struktur tiga dimensi protein spike receptor binding domain hamper identic dengan *SARS-CoV*. Dengan seperti itu juga menemukan bahwa *SARS-CoV-2* tidak digunakan reseptor virus korona lainnya seperti *Aminopeptidase N* (APN) dan *Dipeptidyl peptidase-4* (DPP-4) (Susilo et al., 2020).



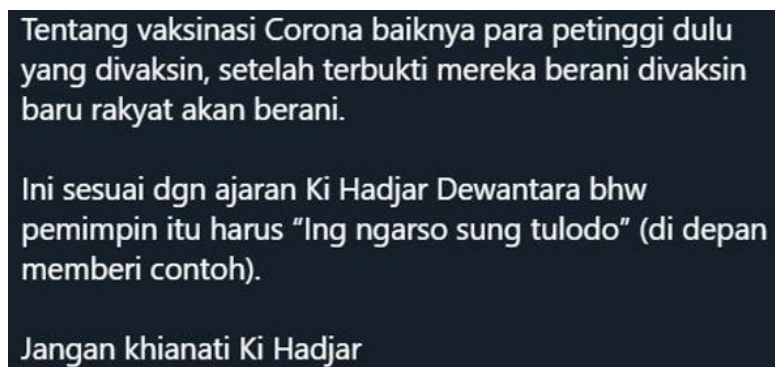
### 2.1.2. Vaksin

Vaksin Sinovac telah diproduksi menggunakan metode yang baik untuk mematikan virus *Covid-19* sehingga vaksin yang dihasilkan tidak mengandung senyawa virus hidup maupun virus yang dilemahkan (Rahayu, 2021). Dilihat dari laman *WHO* atau *World Health Organization*, vaksin corona yang telah dikembangkan sampai saat ini mengandung antigen yang sama dengan antigen yang menyebabkan penyakit. Namun ternyata antigen yang ada di dalam vaksin corona tersebut dikendalikan atau bisa dibilang dilemahkan yang menyebabkan orang yang di vaksin menjadi hilang akan virus corona tersebut (Emanuel et al., 2021). Penelitian sebelumnya telah menunjukkan bahwa keraguan vaksin adalah fenomena umum secara *global*, dengan variabilitas dalam alasan yang dikutip di balik penolakan penerimaan vaksin. Itu alasan paling umum termasuk: risiko yang dirasakan dan manfaat, keyakinan agama tertentu dan kurangnya pengetahuan dan kesadaran. Alasan yang disebutkan di atas dapat diterapkan Keragu-raguan vaksin *COVID-19*, seperti yang ditunjukkan publikasi terbaru menunjukkan kuatnya korelasi antara niat untuk mendapatkan vaksin virus corona dan persepsi keamanannya sikap negatif terhadap vaksin *COVID-19* dan keengganan untuk mendapatkan vaksin. Analisis faktor-faktor tersebut diperlukan untuk mengatasi Keragu-raguan vaksin *COVID-19*, menyusul penilaian cakupan dan besarnya ini ancaman kesehatan masyarakat. Ini dapat membantu dalam memandu tindakan intervensi yang ditujukan membangun dan memelihara tanggapan untuk mengatasi ancaman ini (Sallam, 2021).

### 2.1.3. Twitter

Ekosistem media sosial, yang dikembangkan oleh platform online seperti *Twitter*, menyediakan lingkungan tempat beraneka ragam individu dapat dengan mudah berbagi, berdiskusi, dan terlibat ilmu. Satu studi tahun 2017 melaporkan bahwa 1% -5% dari 187 *Twitter* juta pengguna adalah ilmuwan aktif (Cormier & Cushman, 2021). Dari sudut pandang ilmuwan individu, satu manfaat dari sebuah kehadiran online aktif, terutama di *Twitter*, adalah bantuan di penyebaran pekerjaan. Pengikut akan melihat *tweet* dan algoritma *Twitter* meningkatkan visibilitas lebih lanjut. Oleh karena itu, mudah untuk memahami bagaimana *tweet* yang bisa menjadi *timeline* ilmuwan di berbagai bidang (Ke et al., 2016). Di Salah satu penelitian, seseorang yang meneliti di *Twitter* cenderung beragam bidang. Dengan meningkatnya seruan untuk melibatkan pasien dan pihak nonakademik lainnya dalam perencanaan penelitian ilmiah, ini menjadi semakin penting untuk dipertimbangkan Jangkauan sosial media juga berkontribusi pada penyebaran pengetahuan yang cepat dunia terkait munculnya penyakit virus *Covid-19* (Makris, 2020). Distribusi penelitian yang lebih luas juga dapat meningkatkan dampak positif atau

negatif. Inilah sebabnya mengapa para ilmuwan berusaha keras untuk menerbitkannya di jurnal-jurnal terkenal. *Twitter* dapat semakin memperkuat jangkauan dan pengaruh seseorang (Chi & Cushman, 2020). Dengan seperti ini *Twitter* dapat menjadi salah satu media untuk melakukan penelitian terkait analisis sentimen. Cara mendapatkan data *twitter* tersebut dilakukan dengan cara *Crawling* atau *Scraping* menggunakan *Selenium* pada Bahasa *Python*. Fokusnya adalah pada kata kunci, bahasa, dan lokasi pengguna akun (Trajkova et al., 2020). Contoh *tweet* pada *twitter* dilihat dari gambar 2.1 dibawah ini.



Gambar 2.1. Contoh Tweet Pada Twitter

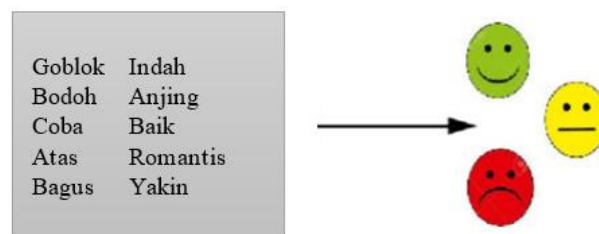
#### 2.1.4. Analisis Sentimen

Analisis sentimen menggunakan pemrosesan bahasa alami (*NLP*), analisis teks, dan teknik komputasi untuk mengotomatiskan ekstraksi atau klasifikasi sentimen dari tinjauan sentimen. Analisis sentimen dan opini ini telah tersebar di banyak bidang seperti informasi konsumen, pemasaran, buku, aplikasi, situs web, dan sosial (Hussein, 2018). Diperlukan pendekatan pemrograman untuk analisis sentimen dengan menangani banyak submasalah terlibat dalam mengekstraksi makna dan polaritas dari teks (Cambria et al., 2017). Analisis sentimen atau penggalan opini adalah studi komputasi opini, sentimen serta emosi. Awal dan pertumbuhan yang cepat dari bidang ini bertepatan dengan media sosial di Web, misalnya, ulasan, diskusi forum, blog, mikro blog, *Twitter*, dan jejaring sosial, karena untuk pertama kalinya dalam sejarah manusia. Itu juga banyak dipelajari dalam penambangan data, penambangan web, penambangan teks, dan pengambilan informasi. Nyatanya, memang demikian menyebar dari ilmu komputer ke ilmu manajemen dan ilmu sosial seperti pemasaran, keuangan, ilmu politik, komunikasi, ilmu kesehatan, bahkan sejarah, karena kepentingannya untuk bisnis dan masyarakat secara keseluruhan. Perkembangan ini disebabkan oleh fakta bahwa pendapat adalah pusatnya hampir semua aktivitas manusia dan merupakan pengaruh utama dari perilaku kita. Keyakinan dan persepsi tentang realitas, dan pilihan yang kita buat, pada tingkat tertentu,

dikondisikan pada bagaimana orang lain melihat dan mengevaluasi dunia. Untuk alasan ini, kapan pun kita perlu membuat keputusan, kita sering mencari pendapat orang lain. Ini tidak hanya berlaku untuk individu tetapi juga untuk organisasi. Oleh karena itu, analisis sentimen secara otomatis seperti ini sangat diperlukan (Zhang et al., 2018).

### 2.1.5. Metode *Lexicon Based*

Metode *Lexicon Based* adalah cara mengelompokkan kata-kata ke dalam kelompok sentimen positif dan sentimen negatif. Misalnya kata-kata seperti kategori "cantik", "baik", "pintar" dalam sekelompok kata yang memiliki sentimen positif, kemudian kata-kata seperti kategori "buruk", "buruk", "bodoh" dalam sekelompok kata yang memiliki sentimen negatif. Keberadaan sebuah kata dalam kelompok polaritas sentimental merepresentasikan implikasi emosi yang terkandung dalam kata tersebut. Polaritas sentimen suatu kata dalam kelompok sentimen dinyatakan dengan nilai yang menyatakan bobot hubungan kata tersebut dengan kelompok sentimennya. Jadi, leksikon sentimen adalah sumber leksikal yang berisi informasi tentang emosi yang terkandung dalam beberapa kata (Christina & Ronaldo, 2020). Leksikon berisi daftar kata atau frase dan ini adalah sumber daya penting dalam analisis sentimen. Ada beberapa pendekatan yang benar membangun leksikon secara manual atau otomatis. Leksikon manual adalah waktu yang mahal dan tidak bekerja dengan semua domain sehingga *Lexicon* otomatis menjadi topik penelitian yang lagi hangat karena mudah digunakan dan dikerjakan domain apa saja (Abd et al., 2021). Metode berbasis leksikon menghitung polaritas sentimen sebagai fungsi dari kata-kata yang memiliki sentimen melalui media sosial *twitter* (Bagheri & Islam, 2017). *Lexicon based* dilakukan dengan menggunakan rumus untuk klasifikasi sentimen yaitu *SentiWordNet* sentimen yaitu *SentiWordNet* (Kusumawati, 2017). Alur metode klasifikasi *lexicon based* dapat dilihat dari gambar 3 dibawah ini.



**Gambar 2.2. Contoh Klasifikasi Lexicon Based**

Dilihat dari Gambar 2.2 alur metode klasifikasi *lexicon based* untuk menganalisis sentimen ialah dengan mengklasifikasikan sebuah kata-kata yang ada apakah dikategorikan positif ataupun negatif (Hamdan, 2016). Dengan seperti itu maka penelitian ini akan merujuk pada

dataset yang telah di ambil dari media sosial *Twitter* kemudian di klasifikasikan apakah *Tweet* yang membahas tentang vaksin lebih cenderung positif atau negatif.

Dalam proses klasifikasi sentimen menggunakan metode *lexicon based* dilakukan pada setiap kata yang ada pada kalimat di dataset dengan *SentiWordNet*. Tetapi untuk penelitian ini akan menggunakan dataset *lexicon* Bahasa Indonesia Pemilihan kata yang memiliki lebih dari satu arti maka synset akan dilakukan berdasarkan metode *First Sense* dari *SentiWordNet* tersebut dengan memperhatikan mana yang muncul paling atas atau yang lebih populer. Setelah itu, kata yang berhasil di klasifikasi sesuai yang ada di *SentiWordNet* kemudian untuk mendukung proses mencari nilai sentimen menggunakan rumus sebagai berikut :

$$S_{positive} = \sum_{i \in t}^n positive\ score_i \quad (1)$$

$$S_{negative} = \sum_{i \in t}^n negative\ score_i \quad (2)$$

Dimana ( $S_{positive}$ ) adalah suatu nilai dari kalimat yang didapatkan dari penjumlahan n skor polaritas kata opini positif dan ( $S_{negative}$ ) adalah suatu nilai dari kalimat yang didapatkan dari penjumlahan n skor polaritas kata opini negatif. Kemduain nilai pada setiap kata yang ada di kalimat digunakan sebagai acuan untuk melihat adanya perbandingan antar kata tersebut.. Sehingga dalam satu kalimat akan diketahui total jumlah nilai positif ( $S_{positive}$ ) dan juga nilai negatif ( $S_{negative}$ ) dari tiap-tiap kata penyusunnya. Dari persamaan suatu nilai sentimen yang ada dapat ditentukan persamaan 3 untuk menentukan orientasi sentimen dengan perbandingan yang ada pada jumlah positif dan negatif.

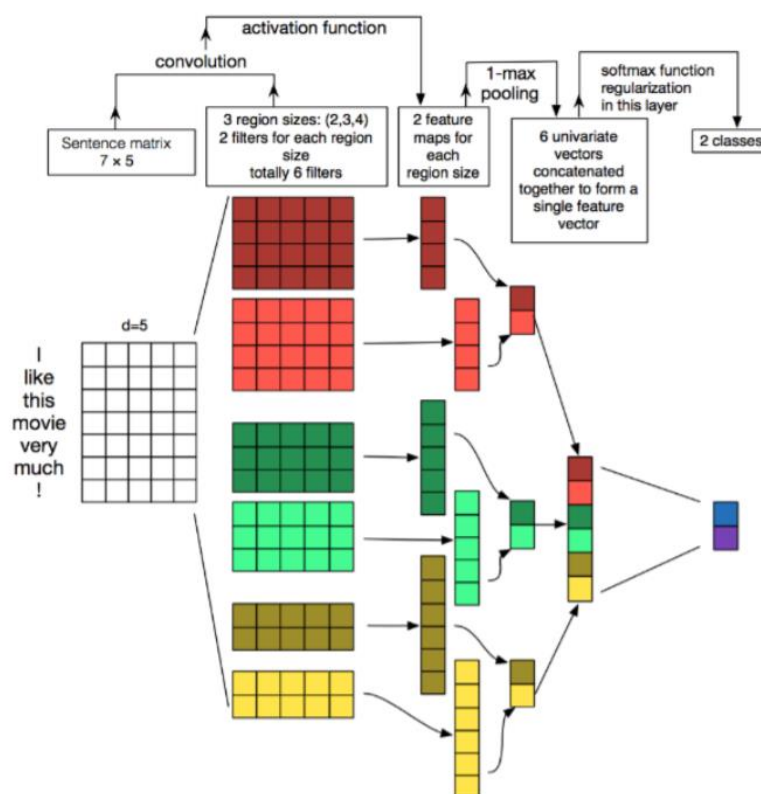
$$Sentence_{sentiment} = \begin{cases} positive & \text{if } S_{positive} > S_{negative} \\ neutral & \text{if } S_{positive} = S_{negative} \\ negative & \text{if } S_{positive} < S_{negative} \end{cases} \quad (3)$$

Dari rumus diatas dapat ditentukan bahwa jika total jumlah nilai positif lebih besar dari jumlah nilai negatif maka kalimat memiliki sentimen positif. jika negatif maka sebaliknya.

### 2.1.6 Convolutional Neural Network

*Convolutional Neural Network* (*ConVet / CNN*) adalah algoritma pembelajaran mendalam yang digunakan oleh untuk mengklasifikasikan gambar. *CNN* umumnya digunakan untuk mendeteksi dan mengenali objek dalam gambar. *CNN* pada dasarnya adalah lapisan konvolusi, yang menerapkan fungsi pemicu non-linear (seperti *ReLU* atau *tanh*) untuk menghasilkan model. Pada tahap *training*, *CNN* akan secara otomatis mempelajari nilai filter berdasarkan

hasil yang diinginkan. *CNN* juga dapat digunakan untuk pemrosesan bahasa alami atau *NLP* dengan mengubah representasi kalimat menjadi array. *CNN* secara sederhana menggunakan beberapa vektor statis dan pengaturan *hyperparameter* bisa mendapatkan hasil yang sangat baik melalui beberapa eksperimen yang berbeda. Ye Zhang menyinggung arsitektur *CNN* dalam penelitiannya yang berjudul "*A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional*", yang memiliki 3 jenis ukuran area filter yang masing-masing memiliki 2 filter (Khatami, et al.,2020). Berikut adalah contoh arsitektur *CNN* :



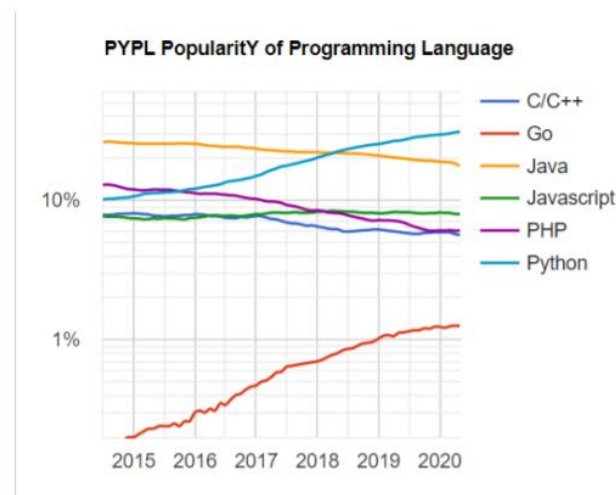
**Gambar 2.3. Arsitektur CNN**

Di CNN untuk distribusi NLP, inputnya adalah kalimat atau dokumen yang direpresentasikan sebagai array. Setiap baris dalam matriks adalah simbol, biasanya kata atau karakter dalam bentuk vektor. Pada penjelasan Gambar 2.3 di atas, kalimat yang digunakan sebagai masukan adalah "Saya sangat menyukai film ini!" dan matriks 7x5 dibuat. Kemudian dijelaskan 3 area filter dengan ukuran 2, 3 dan 4 dan masing-masing area filter memiliki 2 filter. Setiap filter menggulung matriks kalimat dan menghasilkan peta fitur. Kemudian dilakukan pengelompokan 1max pada setiap kartu dan dicatat jumlah hasil pengelompokan terbesar dari setiap kartu. Oleh karena itu, vektor fitur non-variabel dihasilkan dari 6 kartu dan kemudian

digabungkan untuk membentuk vektor fitur untuk lapisan terakhir. Lapisan softmax terakhir menerima vektor fitur sebagai input dan menggunakannya untuk klasifikasi kalimat (Wianto, PWA 2018).

### 2.1.7. *Python*

*Python* saat ini adalah bahasa pemrograman yang tumbuh paling cepat di dunia, berkat itu kemudahan penggunaan, pembelajaran cepat, dan berbagai modul yang berkualitas tinggi untuk ilmu data dan pembelajaran mesin. Namun yang mengejutkan, *Python* sangat mudah digunakan daripada pemrograman. *Python* juga sangat mendukung *NLP* atau *natural language processing* (Qi et al., 2020). Dengan fokus utamanya pada *readability*, *Python* adalah bahasa pemrograman yang diinterpretasikan tingkat tinggi, yang dikenal luas karena mudah dipelajari, namun tetap dapat memanfaatkan kekuatan tingkat sistem Bahasa pemrograman bila perlu. Selain dari manfaat bahasanya itu sendiri, alat dan pustaka yang tersedia membuat *Python* sangat menarik untuk beban kerja dalam data sains, pembelajaran mesin, dan komputasi ilmiah. Menurut jajak pendapat *KDnuggets* baru-baru ini menyurvei lebih dari 1800 peserta untuk mengetahui preferensi dalam analitik, ilmu data, dan pembelajaran mesin, *Python* mempertahankan posisinya di puncak bahasa yang paling banyak digunakan pada tahun 2019 (Raschka et al., 2020). Dalam penelitian ini, *tweet* diambil dari data sosial *Twitter* yang disebut *tweet* sesuai dua kata kunci pencarian tertentu yaitu *#vaksin* dan *#covid-19* untuk mengekstrak *tweet* dan melakukan analisis sentimen pada dataset dan bahasa pemrograman *Python* telah terpilih. *Python* menyediakan banyak pustaka yang mudah digunakan untuk mengakses *platform* media sosial *Twitter*. *Python* dapat mengakses *tweet* ini dari *API* pencarian *Twitter* dan perpustakaan *tweepy*. Singkatnya, Pendekatan analisis sentimen telah diterapkan pada data yang telah di kumpulkan (H. Manguri et al., 2020). Semua data dikumpulkan dengan menggunakan *tweepy library* dan analisis sentimen dilakukan dengan menggunakan pustaka *TextBlob Python* di *Google Colab* (Pokharel, 2020). Informasi tersebut kemudian disimpan ke dalam file *CSV* untuk diproses kemudian. Pustaka yang digunakan termasuk *Pandas*, *Matplotlib*, dan *Notebook Jupyter* atau *Google Colab* (Gunnarsson & Herber, 2020). Untuk lebih mendukung mengapa *Python* sangat populer dibandingkan Bahasa pemrograman lainnya, perhatikan gambar dibawah ini.



**Gambar 2.4. Most Popular Programming Language**

Pada Gambar 4 bisa dilihat dari tahun 2015-2020 bahwa perubahan grafik tingkat popularitas Bahasa pemrograman *Python* masih cenderung naik di bandingkan pesaing lainnya yaitu *Go*, *Java*, *Javascript* dan *PHP* (Gunnarsson & Herber, 2020).

## 2.2. Tinjauan Pustaka

Pada tinjauan pustaka, penulis mereferensikan beberapa penelitian sebelumnya yang serupa dengan judul Internship II di laporan ini, diantara lain ada pada tabel berikut ini :

**Tabel 1. Perbandingan Penelitian Sebelumnya Dengan Penelitian Yang Akan Dilakukan**

No	Judul	Penulis	Hasil	Pengembangan
1	Analisis Sentimen Pro dan Kontra Masyarakat Indonesia tentang Vaksin <i>COVID-19</i> pada Media Sosial <i>Twitter</i>	Fajar Fathur Rachman, Setia Pramana Politeknik Statistika STIS, Jakarta, Indonesia	Berdasarkan yang penulis baca hasil dari analisis sentimen yang dapat ditarik dari penelitian tersebut bahwa masyarakat lebih banyak memberikan respon yang positif terhadap vaksin <i>COVID-19</i> dibandingkan dengan respon yang negatif. Proses klasifikasi pada	Pengembangan yang akan penulis lakukan untuk penelitian ini adalah penulis akan mengklasifikasikan berdasarkan daerah di Indonesia. Misalnya kota Medan apakah respon positif atau negatif. Dengan seperti itu kita dapat mengetahui daerah mana saja di Indonesia

			penelitian ini menggunakan metode <i>LDA</i> .	yang responnya positif atau negatif.
2	<i>Studying Public Perception about Vaccination: A Sentiment Analysis of Tweets</i>	Viju Raghupathi, Jie Ren and Wullianallur Raghupathi New York	Hasil pada penelitian ini adalah mengklasifikasikan analisis sentimen terkait vaksin dari media sosial <i>Twitter</i> menggunakan <i>Teknik TF-IDF</i> .	Pengembangan yang akan penulis lakukan ialah melakukan metode lain untuk proses analisis sentimen yaitu dengan menggunakan metode <i>Lexicon Based</i>
3	<i>Sentiment analysis of social media posts on pharmacotherapy: A scoping review</i>	Chanakya Sharma, Samuel Whittle, Pari D. Haghighi, Frada Burstein Australia	Hasilnya ialah analisis sentimen lebih cenderung mengarah ke respon masyarakat terkait adanya obat-obatan tertentu yang ada pada vaksin. Metode yang dipakai adalah <i>lexicon based</i> dan <i>machine learning</i> .	Pengembangan yang penulis lakukan ialah dengan melakukan ulang <i>scraping dataset</i> pada media sosial <i>Twitter</i> agar data yang diperoleh lebih baru.
4	<i>Sentiment analysis of tweets about COVID-19 disease during pandemic</i>	Goran Matošević, Vanja Bevanda Pula, Hrvatska	Hasil yang didapat pada penelitian ini ialah dengan mengklasifikasikan sentimen di berbagai macam negara yaitu USA, UK, Italy, Spain, Germany dan Sweden	Pengembangan yang penulis lakukan setelah membaca penelitian ini adalah dengan klasifikasi top 10 daerah atau kota yang ada di Indonesia
5	Vaksin Covid 19 Di Indonesia :	Rochani Nani	Hasilnya adalah sebuah Analisa dari data media sosial <i>Twitter</i> tentang	Pengembangan yang penulis lakukan yaitu dengan menggunakan



	Analisis Berita Hoax	Rahayu, Sensusiyati  Pusat Data Dan Dokumentasi Ilmiah LIPI	berita hoax terkait vaksin.	hasil Analisa berita hoax tersebut dilakukan penulis mendapatkan ide untuk membuat analisis sentimen positif atau negatif respon masyarakat terhadap vaksin <i>covid-19</i> .
--	-------------------------	---	--------------------------------	---

## BAB III

### OBJEK STUDI

#### 3.1. Material

Material adalah sebuah data yang berfungsi untuk diolah dari suatu penelitian (Mendez Palencia, C. 2021). Data yang penulis ambil untuk mengerjakan penelitian ini tentang analisis sentimen terhadap vaksin covid-19 ialah menggunakan data dari media sosial twitter. Kemudian, cara untuk mendapatkan atau mengumpulkan data twitter tersebut ialah dengan menggunakan Teknik scraping data. *Scraping* adalah suatu Teknik untuk mengambil dan mengumpulkan data dari suatu website atau media sosial dengan semi-terstruktur (Setiawan, et al 2020). Disini penulis menggunakan media sosial twitter dengan menggunakan kata kunci vaksin covid-19 serta lokasi berdasarkan provinsi di Indonesia. Pada media sosial twitter akan diambil beberapa atribut yang digunakan untuk penelitian ini yaitu *username*, *date*, *content*, *user\_location* dan bulan.

**Tabel 3.1. Atribut yang digunakan pada data twitter**

No	Atribut	Keterangan
1	<i>Username</i>	Username dari pemilik akun atau pengguna twitter.
2	<i>Date</i>	Date adalah sebuah tanggal, bulan dan tahun pada saat tweet atau cuitan itu dibuat
3	<i>Content</i>	Content adalah sebuah status atau cuitan yang dibagikan oleh pengguna twitter
4	<i>User_location</i>	User_location ialah lokasi dari pengguna twitter tersebut.
5	Bulan	Bulan adalah pada bulan apa cuitan twitter tersebut di bagikan.

Pada penelitian ini, data yang digunakan menggunakan *library snsrape* dan juga *library tweepy*. Secara singkat *snsrape* akan membantu kita untuk mengumpulkan data tanpa API dan kelebihan nya ialah bisa filter lokasi. Kemudian, *tweepy* adalah *library Python* untuk berinteraksi dengan *API* Twitter (Azizul H. 2020). Data yang diambil pada penelitian ini sebanyak 34000 data dengan masing-masing provinsi kurang lebih 1000 data. 34000 data tersebut akan dibagi beberapa bagian yang akan diberi label positif, negatif dan netral :

1. 1000 data akan di label manual.
2. 5000 data akan di label otomatis menggunakan metode *lexicon based*.
3. 28000 data akan di label otomatis menggunakan metode *CNN*.

Adapun contoh data yang digunakan yaitu :

**Tabel 3.2. Contoh data yang diambil dari media sosial twitter**

No	Username	Date	Content	User_location	Bulan
1	bellzia	2021-02-02 11:16:55+00:00	Hari ini vaksin dosis kedua (booster), gak kerasa ngantuk seperti dosis pertama kemarin, dan gak pegel 😊	Nusa Tenggara Barat	Februari
2	Andreebagaas	2021-01-14 17:13:42+00:00	Aku ndamau di suntik vaksin, vaksin ku adlh dzikir	Jawa Tengah	Januari
3	whitehorsegroup	2021-03-29 11:09:25+00:00	Seluruh driver & crew White Horse Bali sudah mendapat vaksin Covid-19.	DKI Jakarta	Maret
4	niczous	2021-04-28 13:52:36+00:00	Peluang untuk dapat vaksin awal! 🤔 <a href="https://t.co/HyXQwDfk08">https://t.co/HyXQwDfk08</a>	Aceh	April
5	dhido_blenk	2021-07-06 20:52:23+00:00	Buat teman-teman yang hendak berwisata ke Bali, jadi makin yakin kan buat naik Bus White Horse? 😊🙏 <a href="https://t.co/N1kZ6bl5de">https://t.co/N1kZ6bl5de</a>	Nusa Tenggara Timur	Januari

Dari tabel 3.2 diatas yang akan di menjadi kolom utama yang digunakan yaitu kolom *content*, karena akan menghasilkan suatu analisis sentimen positif, negatif atau netral. Kemudian dua kolom lagi yang dibutuhkan yaitu *user\_location* dan *bulan*.

### 3.2. Teknologi

Teknologi adalah suatu hal yang selalu ada untuk menyelesaikan suatu masalah dalam penelitian. Munculnya teknologi baru juga menawarkan cara baru bagi para peneliti untuk mengumpulkan dan menganalisis data (Greenhalgh, S. P., et al. 2020). Adapun beberapa teknologi yang dipakai untuk menyelesaikan penelitian ini adalah :

#### 1. Scraping

*Scraping* atau proses pengambilan data ialah sebuah teknik untuk mengumpulkan data dari *website*. Teknik itu dialokasikan ke dalam tiga fragmen yaitu *scraper web* untuk menarik tautan yang diinginkan dari web, kemudian data diekstraksi untuk mendapatkan data dari tautan sumber dan akhirnya menyimpan data itu ke dalam file csv. Bahasa Python diimplementasikan untuk proses ini. Dengan demikian, karena

*library* pada Python memiliki fungsi yang paling tepat untuk mengambil data yang diinginkan dari situs web yang diinginkan (Thomas, D. M., & Mathur, S. 2019).

## 2. Lexicon Based

Metode *lexicon based* atau berbasis leksikon menggunakan kamus stok kata dengan kata-kata opini dan mencocokkan rangkaian kata yang diberikan dalam sebuah teks untuk menemukan polaritas. Berbeda dengan metode pembelajaran mesin, pendekatan ini tidak perlu melakukan preprocess data tidak harus melatih classifier (Taj, S., Shaikh, B. B., et al. 2019). Metode *lexicon based* pada penelitian ini akan digunakan untuk membuat label sentimen otomatis untuk 6000 data sebagai data training.

## 3. Convolutional Neural Network

*Convolutional neural network* atau jaringan saraf convolutional didefinisikan sebagai algoritma yang biasanya digunakan untuk memproses gambar dan teks data yang termasuk dalam kategori *neural network* algoritma. Kata konvolusi itu sendiri didefinisikan sebagai matriks yang berfungsi untuk mengklasifikasikan dan memfilter gambar dan teks. Ada beberapa lapisan yang digunakan dalam *Convolutional Neural Network* yang berfungsi sebagai filter pada masing-masing proses yang disebut proses pelatihan. Dalam proses pelatihan ada 3 tahapan, yaitu, *convolutional layer*, *pooling layer* and *fully connected layer*. Dari setiap lapisan yang terdapat dalam *convolutional neural* arsitektur jaringan, neuron akan dihubungkan ke yang berikutnya lapisan. *Layer* terakhir akan menampilkan output berupa diklasifikasikan data dari lapisan sebelumnya terhubung (Djati, U. S. G. 2021).

$$L_C = - \sum_{i=1}^N q_i \log(p_i) - \sum_{i=1}^N (1 - q_i) \log(1 - p_i) \quad (4)$$

Di mana N mewakili jumlah total *voxel* dalam output dan  $p_i$  dan  $q_i$  menunjukkan probabilitas bahwa *voxel* ke-i dari kebenaran dasar sesuai dengan masing-masing daerah layer dari hasil segmentasi jaringan (Lee, D. K. 2020).

## 4. Deep Learning

*Deep learning* atau pembelajaran mendalam adalah aplikasi jaringan saraf tiruan yang meniru cara kerja korteks manusia yang memiliki banyak lapisan tersembunyi dan termasuk dalam penelitian machine learning yang berfungsi untuk memberikan akurasi dalam berbagai penelitian seperti mendeteksi suatu objek. Pembelajaran mendalam dapat secara otomatis memproses data seperti gambar atau teks tanpa harus mengenali input dari manusia, sebaliknya ke Pembelajaran Mesin tradisional yang harus dikenali

terlebih dahulu. Algoritma pembelajaran mendalam memiliki fitur yang dapat mengekstrak otomatis proses pemecahan masalah. Algoritma seperti ini sangat diperlukan dalam kecerdasan buatan karena mereka dapat mengurangi beban program dalam memproses masalah. Memecahkan masalah yang dilakukan oleh *deep learning* pada sistem komputer menerapkan konsep hirarki. Konsep hierarkis ini menggabungkan yang sederhana konsep dalam mempelajari konsep yang lebih kompleks. Pembelajaran mendalam dapat belajar dari fungsi pemetaan yang kompleks, dari input hingga output, tidak memerlukan konsep atau input buatan dari manusia (Djati, U. S. G. 2021).

#### 5. *Text Preprocessing*

*Text preprocessing* adalah salah satu langkah dalam *text mining* yang menerima informasi dengan data teks yang tidak sempurna struktur. *Text preprocessing* dalam teks, akan ada beberapa proses yang menciptakan informasi yang sebelumnya dimiliki previously struktur data teks yang tidak sempurna diekstraksi menjadi informasi informasi yang berguna memiliki struktur data yang lebih sempurna dari sebelumnya. Setelah melalui proses *stopword*, kata-kata yang telah dipilih akan diproses untuk menghilangkan imbuhan dalam proses mengubah kata, kata-kata yang dihapus akan kembali ke bentuk kata dasar tanpa mengubah kelompok kata.

#### 6. *Confusion Matrix*

*Confusion matrix* adalah metode perhitungan untuk proses klasifikasi dalam konsep *data mining* untuk menemukan bagaimana data dapat diklasifikasikan dengan benar. Dalam klasifikasi matriks konfigurasi, empat istilah dikenal untuk hasil klasifikasi matriks konfigurasi (Djati, U. S. G. 2021). termasuk:

- a. *TP (True Positive)* adalah data yang terdeteksi dengan benar dan adalah positif.
- b. *TN (True Negative)* adalah data yang terdeteksi dengan benar tetapi adalah negatif.
- c. *FP (False Positive)* adalah data yang terdeteksi salah tetapi adalah positif.
- d. *FN (False Negative)* adalah data yang terdeteksi salah dan adalah negatif.

Secara umum, perhitungan klasifikasi termasuk dalam matriks konfigurasi terdiri dari *recall*, *presisi*, *F1-score*, *accuracy*, dan *support*. *Recall* adalah perhitungan deskripsi kesuksesan tingkat sistem saat memulihkan informasi. *Recall* memiliki rumus persamaan seperti pada persamaan (5) di bawah ini:

$$recall = \frac{TP}{TP + FN} \quad (5)$$

Presisi adalah tingkat akurasi antara jawaban ditampilkan oleh sistem dan informasi yang diharapkan oleh pengguna sistem. Presisi memiliki rumus persamaan seperti pada persamaan (6) di bawah ini:

$$precision = \frac{TP}{TP + FP} \quad (6)$$

*F1-score* atau *F-measure* adalah perhitungan yang digunakan sebagai evaluasi klasifikasi presisi gabungan dan mengingat. Ketika kasus presisi dan *recall* memiliki perbedaan nilai, F1-skor menjadi nilai timbal balik antara presisi dan recall dengan nilai harmonik tertimbang dari presisi rata-rata dan recall. Skor F1 memiliki rumus persamaan seperti yang ditunjukkan pada persamaan (7) di bawah ini:

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} \quad (7)$$

Akurasi adalah nilai akurasi dalam suatu klasifikasi proses. Nilai akurasi berasal dari distribusi hasil dari seluruh klasifikasi yang benar dan jumlah data yang telah diklasifikasikan. Akurasi memiliki rumus yang sama seperti pada persamaan (8) di bawah ini:

$$Accuracy = 100\% \times \frac{True\ Classification\ Total}{Classification\ Total} \quad (8)$$

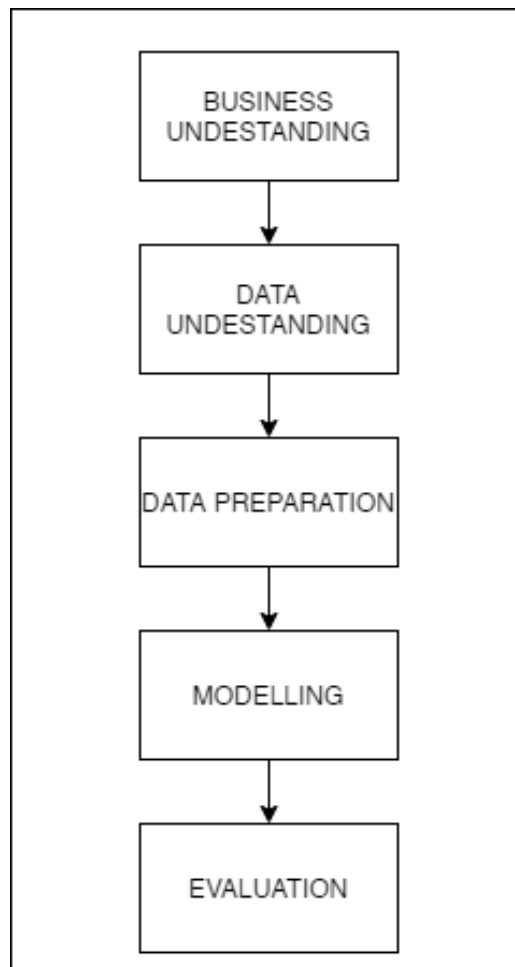
Support adalah nilai total semua dokumen yang diklasifikasikan dengan prediksi klasifikasi benar atau salah. Dukungan bisa digunakan untuk menentukan jumlah dokumen yang muncul.

## BAB IV

### METODOLOGI PENELITIAN

#### 4.1. Diagram Alur Metodologi Penelitian

Metodologi penelitian yang akan dipakai untuk proses penelitian ini penulis mengambil referensi dari konsep *CRISP-DM* atau *Cross-Industry Standard Process for Data Mining*. CRISP-DM (CROSS-Industry Standard Process for Data Mining) merupakan suatu konsorsium perusahaan yang didirikan oleh Komisi Eropa pada tahun 1996 dan telah ditetapkan sebagai proses standar dalam data mining yang dapat diaplikasikan di berbagai sektor industri. Gambar menjelaskan tentang siklus hidup pengembangan data mining. Adapun tahapan yang penulis jelaskan, sebagai berikut :



**Gambar 4.1. Metodologi Penelitian**

## 4.2. Tahapan – Tahapan Diagram Alur Metodologi Penelitian

### 4.2.1. *Business Understanding*

*Business understanding* atau pemahaman bisnis, berarti memahami tujuan bisnis, menilai situasi, dan menerjemahkan tujuan bisnis menjadi tujuan penambangan data.. Dalam penelitian ini dibutuhkan pengetahuan untuk mendapatkan data twitter yang kemudian data twitter tersebut dibutuhkan untuk menganalisis sentimen terkait vaksin covid-19.

### 4.2.2. *Data Understanding*

Pada tahap ini data dikumpulkan dan diolah, kemudian menganalisis data tersebut dan mengevaluasi data yang digunakan dalam penelitian ini. Sumber data yang digunakan dalam penelitian ini adalah data dari media sosial twitter.

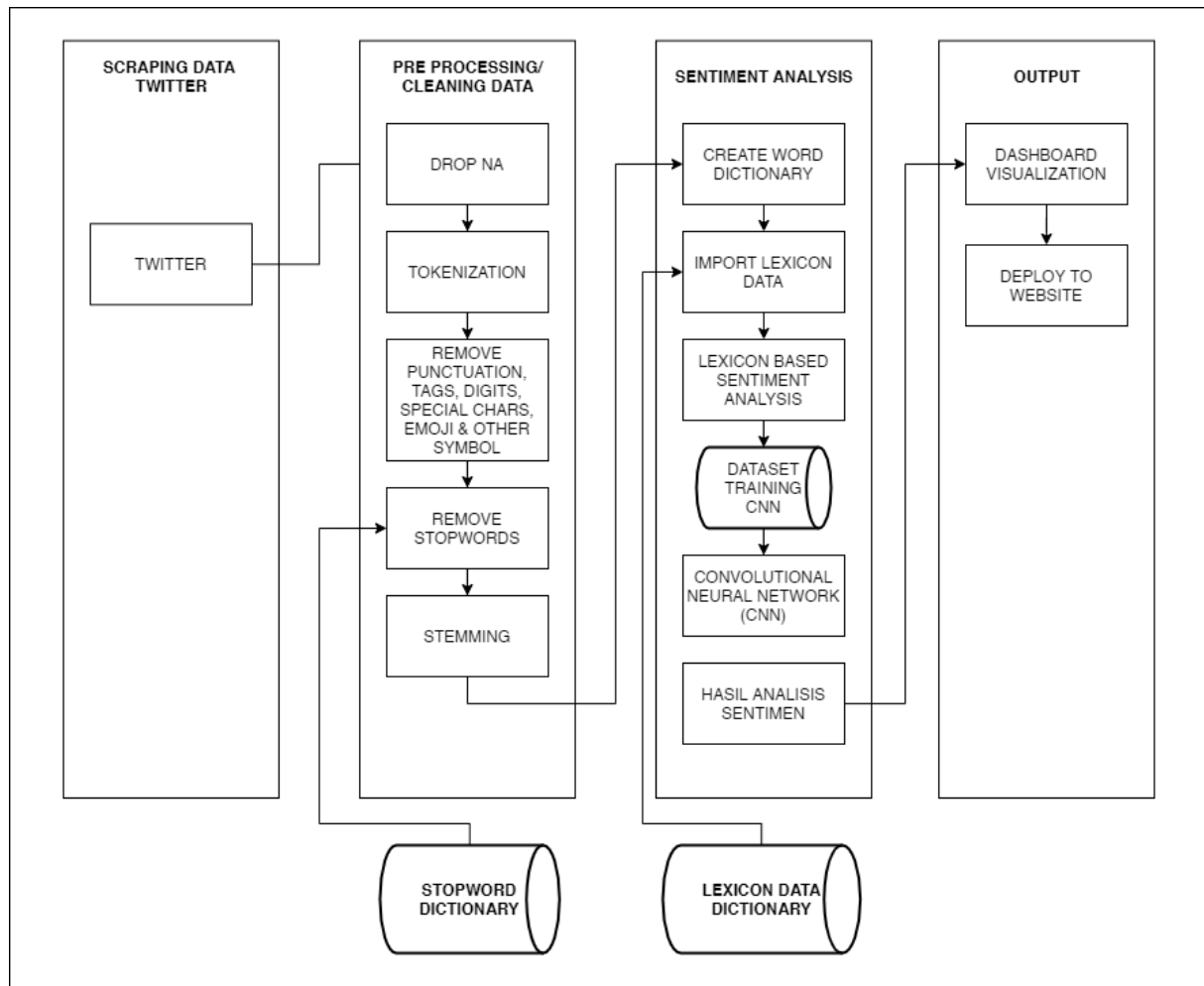
### 4.2.3. *Data Preparation*

*Data Preparation* atau Persiapan Data atau bisa disebut data *preprocessing* adalah suatu proses / langkah-langkah yang dilakukan untuk menjadikan data mentah menjadi data yang berkualitas (input yang baik untuk alat data mining). Dengan demikian proses penelitian akan lebih mudah.

### 4.2.4. *Modelling*

Dalam tahap ini, berbagai macam metode pemodelan dipilih dan diterapkan ke dataset yang sudah disiapkan untuk mengatasi kebutuhan bisnis yang sesuai. Adapun metode yang digunakan yaitu *Lexicon Based* dan CNN atau *Convolutional Neural Network*, Teknik yang dipakai ialah *semi-supervised learning*. Teknik *semi-supervised learning* adalah jenis pembelajaran mesin di mana suatu kasus memiliki sejumlah besar data input dan hanya beberapa data yang diberi label. Masalah ini terletak antara *supervised learning* dan *unsupervised learning*. Adapun sesuai dengan metodologi penelitian yang dijelaskan bahwa penulis terlebih dahulu memberi label beberapa data dan kemudian akan menggunakan *Lexicon Based* untuk melakukan labelling secara otomatis, kemudian terakhir menggunakan metode CNN untuk melakukan *modelling* pembuatan program. Metode penelitian lainnya yang dilakukan untuk penelitian ini penulis akan melakukan proses implementasi seperti, *Crawling Data Twitter*, *Pre Processing* atau *Cleaning Data*, *Proses Sentiment Analysis* dan yang terakhir Output berupa deploy yang berupa input teks dan akan menghasilkan keluaran sentimen positif, negatif ataupun netral. Deploy tersebut memakai *framework flask* dari python dan visualisasi menggunakan platform *Tableau*. Penjelasan tersebut penulis jelaskan pada Diagram Alir dibawah ini.





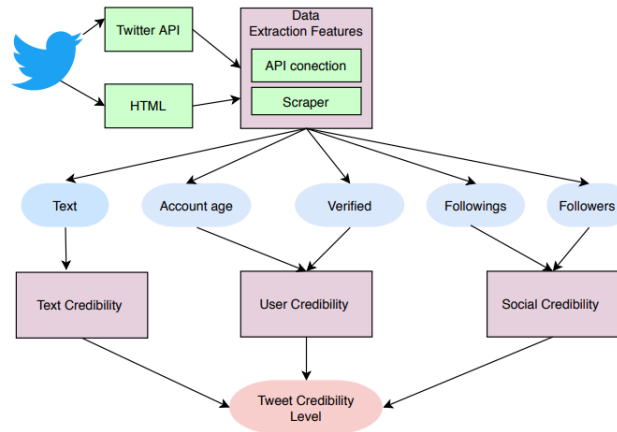
**Gambar 4.2. Diagram Alir Penelitian**

Pada Gambar 4.2 dapat dilihat proses dari awal sampai akhir penelitian yang akan dilakukan pada penelitian ini. Adapun tahap-tahap dalam penelitian ini adalah:

#### 4.2.4.1. Scraping Data

Pada tahap ini dilakukan pengumpulan data dari media sosial *twitter* dengan menggunakan library *snsrape* yang tidak memakai *API* dan juga *tweepy* dengan menggunakan *API*. Seperti *username*, *text tweet*, *location* dll. Ada dua cara mendapatkan data *Twitter* yaitu menggunakan *API* dan manual. Dengan menggunakan *snsrape* menjadi lebih mudah untuk mendapatkan *user\_location* yang dapat ditentukan langsung dengan filter lokasi, tetapi kelemahan dari *snsrape* karena tidak menggunakan *API* jadi terkendala pada limit yang sedikit. Oleh karena itu, penulis menambahkan library *tweepy* yang memakai *API* untuk mendapatkan data yang lebih banyak. Pada penjelasan bab sebelumnya dijelaskan bahwa data yang diambil pada penelitian ini sebanyak 34000 data dengan masing-masing provinsi kurang lebih 1000 data. 34000 data tersebut akan dibagi beberapa bagian sebagai berikut :

1. 1000 data akan di label manual.
2. 5000 data akan di label otomatis menggunakan metode *lexicon based*.
3. 28000 data akan di label otomatis menggunakan metode *CNN*.



**Gambar 4.3. Diagram Alir Model Proses Pengumpulan Data *Twitter***

Untuk implementasi metode ekstraksi *API*, harus meminta izin untuk menggunakan *API Twitter*. Untuk mendapatkan *API*, akun *Twitter* harus membuat request, lalu terdaftar sebagai akun pengembang di *developer.twitter.com* situs. Formulir diisi untuk meminta *API*. Setelah *Twitter* memberikan izin dan *API*-nya diperoleh, data berhasil diekstraksi (Dongo et al., 2020).

#### 4.2.4.2. *Preprocessing/Cleaning Data*

Tahap ini merupakan proses menghilangkan dan membersihkan data yang masih mentah atau lebih mudah diartikan sebagai suatu proses/langkah yang dilakukan untuk membuat data mentah menjadi data yang berkualitas (input yang baik untuk *data mining tools*).

Pemrosesan awal merupakan langkah persiapan data yang diperlukan untuk klasifikasi sentimen. Untuk melakukan *preprocessing* memungkinkan konfigurasi berikut:

##### a. *DROP NA*

Proses membersihkan data yang pertama ialah dengan cara melakukan *drop NA*, dimana *NA* adalah data yang kosong atau salah satu kolom pada dataset tidak ada atau *none*. Di tahap ini data yang kosong akan di *drop* atau di hapus.

b. *Tokenization*

Teknik ini membagi dokumen menjadi kata / istilah, membentuk vektor kata, yang dikenal sebagai *bag-of-word*. Proses *tokenization* adalah membagi teks yang awalnya berupa kalimat nanti akan dipisah dan menjadi token-token.

c. *Remove Punctuation, Tags, Digits, Special Chars, Emoji & Other Symbol*

Pada tahap ini akan dilakukan penghapusan tanda baca yang terdapat pada dataset selain tanda baca sebuah *tag*, angka, spesial karakter, *emoticon* dan lainnya yang biasa ada di data twitter akan dihapus.

d. *Remove Stopwords*

Ini adalah teknik yang menghilangkan kata-kata yang sering digunakan yang tidak berarti dan tidak berguna untuk klasifikasi teks. Ini mengurangi ukuran korpus tanpa kehilangan informasi penting.

e. *Stemming*

*Stemming* bekerja dengan membuang akhiran kata, menurut beberapa tata Bahasa aturan dan mendapatkan kata dasar dari kata tersebut.

#### 4.2.4.3. *Sentiment Analysis*

Pada tahap ini dilakukan proses sentimen analisis untuk melakukan klasifikasi positif dan negatif dari data yang telah di olah sebelumnya. Di tahap ini akan menggunakan *lexicon data dictionary* yang didapatkan dari penelitian-penelitian yang ada sebelumnya dengan seperti itu maka proses untuk melakukan label secara otomatis akan lebih mudah dan kemudian data yang dari *modelling lexicon based* tersebut akan digunakan sebagai data training untuk metode *CNN*. Adapun metode yang dipakai pada proses analisis sentimen ini ialah *Lexicon Based* dan *Convolutional Neural Network* atau yang biasa disingkat dengan *CNN*. Penjelasan terkait alur dari analisis sentimen ialah sebagai berikut :

1. Pertama 1000 data akan di label secara manual dengan 3 kelas yaitu positif, negatif atau netral. Hasil ini akan menjadi perbandingan akurasi untuk metode *lexicon based*.
2. Dilakukan *modelling lexicon based* dengan menggunakan 1000 data yang sudah di labelin secara manual. Dengan seperti ini maka didapat akurasi dengan membandingkan metode *lexicon based* dengan hasil yang secara manual.
3. Hasil dari tahap kedua jika didapatkan akurasi diatas 75% maka sudah termasuk kedalam akurasi yang cukup bagus untuk melakukan label otomatis dengan 5000 data.
4. Dilakukan *modelling lexicon based* dengan 5000 data yang di labelin secara otomatis.

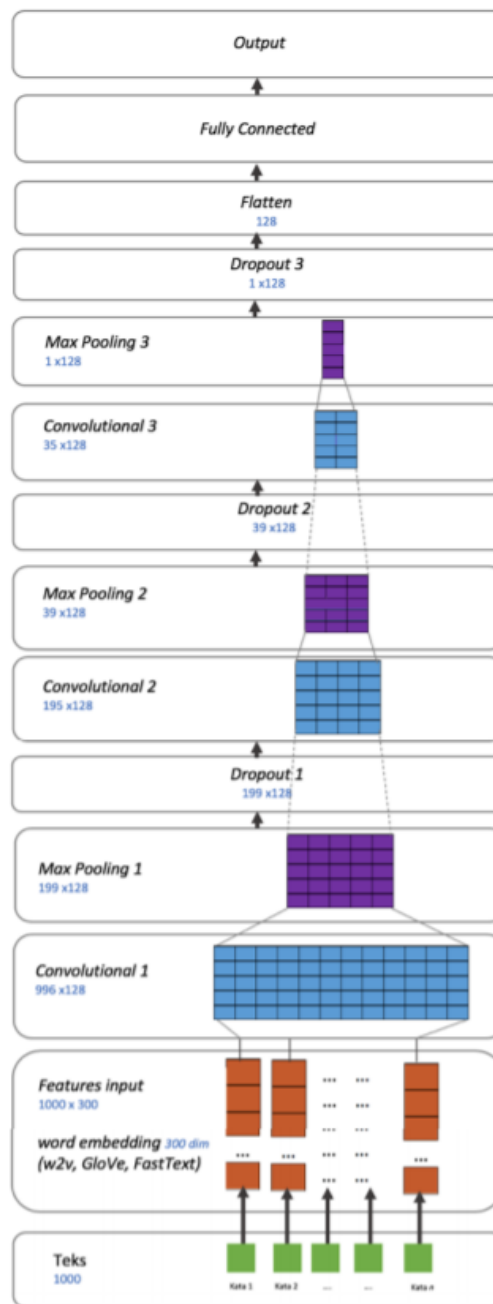
Metode klasifikasi dengan menggunakan *lexicon based* memanfaatkan sebuah kamus leksikon yang berisi daftar kata positif dan negatif berikut dengan nilai polaritasnya. *Lexicon based classification* memanfaatkan metode *rule based method* untuk mengklasifikasi data dengan cara membandingkan jumlah kata positif dan jumlah kata negatif pada satu dokumen. Kamus untuk pendekatan berbasis leksikon dapat dibuat secara manual atau secara otomatis menggunakan kata-kata inti untuk memperluas daftar kata.

5. 1000 data yang dilakukan label secara manual di merge dengan 5000 data yang sudah diberi label secara otomatis menggunakan *lexicon based*. Hasilnya didapat 6000 data untuk dijadikan dataset training *CNN*.
6. Melakukan *modelling CNN* dengan beberapa Teknik yang ada didalamnya yaitu, dengan melakukan perbandingan beberapa model dalam *word embedding* yang penulis gunakan *fasttext*, *word2vec* dan *GloVe*.

*Word2vec* adalah algoritma penyisipan kata yang menetapkan setiap kata dalam teks ke vektor. milik Mikolov. menemukan algoritma *word2vec*. Sejak dimulai pada tahun 2013, model penggabungan kata telah digunakan secara luas dalam penelitian NLP. *Word2vec* merepresentasikan sebuah kata sebagai vektor yang membuat arti dari kata tersebut. Model penyisipan kata ini adalah aplikasi pembelajaran tanpa pengawasan yang menggunakan jaringan saraf tiruan yang terdiri dari lapisan tersembunyi dan lapisan yang terhubung penuh. Bobot setiap lapisan dalam matriks adalah jumlah kata dalam korpus dikalikan dengan jumlah neuron yang tersembunyi di lapisan tersembunyi. Matriks bobot pada lapisan tersembunyi dari model pelatihan digunakan untuk mengubah kata menjadi vektor. Matriks bobot ini seperti tabel pencarian, di mana setiap baris mewakili setiap kata, dan kolom mewakili vektor kata. *Word2vec* didasarkan pada informasi bahasa lokal. Semantik belajar dari kata tertentu dipengaruhi oleh kata-kata di sekitarnya. Model ini menunjukkan kemampuan untuk mempelajari pola bahasa sebagai hubungan linier antara vektor kata. dan *word2vec* hanya mengandalkan informasi kata lokal dengan jendela konteks lokal (CBOW dan Skipgram) Algoritma *GloVe* menggabungkan informasi co-occurrence kata atau statistik global untuk mendapatkan hubungan semantik antar kata dalam korpus. Kami menggunakan metode faktorisasi matriks global yang merepresentasikan ada tidaknya sebuah kata dalam dokumen *GloVe* (Pennington, Socher & Manning, 2014). *Word2vec* sering disebut sebagai neural word embedding karena merupakan model jaringan saraf positif, tetapi *GloVe* adalah model bilinear logaritmik atau model berbasis komputasi sederhana. *GloVe* mempelajari hubungan antar kata dengan menghitung frekuensi

kata yang muncul satu sama lain dalam korpus yang diberikan. Rasio kemungkinan kemunculan kata dapat mengkodekan berbagai bentuk makna dan membantu meningkatkan kinerja masalah analogi kata. Pelatihan model GloVe bertujuan untuk mempelajari vektor-vektor kata sedemikian rupa sehingga hasil kali titik dari kata-kata tersebut sama dengan logaritma probabilitas kata-kata tersebut muncul bersama-sama atau probabilitas kemunculannya bersama. FastText (Bojanowski et al., 2017) merupakan pengembangan tambahan dari word2vec, sebuah metode penyisipan kata. Teknik ini memeriksa ekspresi kata berdasarkan informasi dalam subkata. Setiap kata ditampilkan dalam set karakter ngram. Ini dapat membantu Anda memahami arti dari kata-kata pendek tersebut. Dan di akhir kata untuk menyertakan, buatlah itu dimengerti untuk awalan. Sebuah representasi vektor dikaitkan dengan setiap huruf dalam gram, dan sebuah kata dinyatakan sebagai jumlah dari representasi vektor ini. Setelah sebuah kata direpresentasikan sebagai simbol ngram, model skip gram dapat memeriksa vektor penyisipan kata. Secara umum, model yang merepresentasikan sebuah kata sebagai vektor mengabaikan bentuk kata tersebut, dan setiap kata memiliki vektornya sendiri. Ini adalah batas kata dalam bahasa dengan banyak kata baru dalam kosakata. *FastText* bekerja dengan sangat baik dan dapat memberikan representasi kata yang muncul dalam data pelatihan, memungkinkan Anda melatih model dengan cepat menggunakan kumpulan data besar. Jika Anda tidak melihat kata-kata yang Anda gunakan untuk melatih model Anda, Anda dapat membagi model menjadi ngram dan menyertakan vektornya. ... setelah kata-kata apa yang lebih benar, langkah selanjutnya adalah mengembangkan metode *CNN*.

Di bidang pemrosesan bahasa alami, pembelajaran mendalam telah terbukti dapat diandalkan dalam banyak masalah klasifikasi. Salah satunya adalah Convolutional Neural Network (CNN), yang secara efektif dapat menangkap representasi kalimat yang bermakna, seperti klasifikasi dan pemodelan bahasa (Kalchbrenner, Grefenstette & Blunsom, 2014), analisis sentimen (Kim, 2014) hingga ekstraksi informasi (Chen et al., 2015; Nguyen dan Grishman, 2015; Nurdin dan Maulidevi, 2018). Dalam penelitian ini, CNN satu dimensi digunakan untuk memodelkan peringkat artikel berita pada beberapa topik dalam dataset. CNN satu dimensi ini sangat efektif dalam menurunkan fitur segmen fixed-length dari seluruh kumpulan data, dan sangat cocok untuk masalah pemrosesan bahasa alami (NLP). Tidak ada fitur tambahan yang bersifat handmade, yaitu fitur yang melibatkan pengetahuan ahli bahasa. Algoritme mempelajari semua fitur langsung dari kumpulan data.



**Gambar 4.4. Arsitekur CNN untuk Analisis Setnimen**

Gambar 4.4 menunjukkan arsitektur *Convolutional Neural Network* yang digunakan, terdiri dari *layer input*, *convolutional*, *max pooling*, dan *fully connected*.

- a. Input Layer Teks setiap artikel berita di lapisan input pertama-tama diubah menjadi representasi vektor kata-kata melalui penyisipan kata, dan kemudian dimasukkan ke

lapisan input. Panjang urutan input maksimum adalah 1000, jadi inputnya adalah array 1000 x 300.

- b. *Convolutional Layer* atau lapisan konvolusi. Pada lapisan ini, akan ada 128 filter dengan ukuran kernel 5, bergerak secara vertikal di seluruh matriks input. Setelah melakukan operasi perkalian dan penambahan antara bobot filter dan bobot matriks input, dan menggunakan fungsi aktivasi ReLU untuk operasi nonlinier, diperoleh peta fitur, yang berisi fitur penting dengan dimensi lebih rendah di lapisan tersembunyi pertama. Kemudian masukan peta fitur dari lapisan tersembunyi pertama ke lapisan konvolusi kedua, dan seterusnya. Model CNN dibentuk menggunakan 3 convolutional layer.
- c. *Max Pooling Layer* akan mengambil nilai maksimum 5 elemen dari jendela satu dimensi untuk mendapatkan informasi terpenting dari peta fitur konvolusi.
- d. *Fully Connected Layer Output* dari *hidden layer* sebelumnya, Ubah menjadi vektor dalam bentuk peta fitur dan tautkan ke lapisan keluaran untuk klasifikasi. Pada layer ini digunakan fungsi aktivasi softmax dan fungsi categorical\_crossentropy loss, karena variabel output dari beberapa tipe direpresentasikan oleh one-hot encoding yang terdiri dari angka 0 dan 1. CNN menggunakan batch size = 128 dan epoch = 20 untuk pelatihan. Dan gunakan fungsi optimasi Adam. Untuk menghindari overfitting, untuk setiap lapisan tersembunyi, teknik pengabaian tuning diterapkan (Srivastava et al., 2014) pada tingkat 0,5, yang secara acak akan menonaktifkan 50% neuron selama fase pelatihan.

#### 4.2.4.4. Output

Tahap terakhir ialah hasil apa yang didapatkan pada penelitian ini, yaitu membuat sebuah *dashboard* melalui *Tableau*. Seluruh hasil akan dibuat kedalam *web* yang akan di *hosting* untuk keperluan tertentu.

#### 4.2.5. Evaluation

Pada tahap terakhir ini, model yang sudah dibuat diuji dan dievaluasi keakuratan. Tahap ini mengukur sejauh mana model yang sudah dipilih memenuhi sasaran-sasaran bisnis dan bila demikian, sejauh manakah itu (apakah perlu lebih banyak model untuk dibuat).

## BAB V

### IMPLEMENTASI DAN PENGUJIAN

#### 5.1. Lingkungan Implementasi

Lingkungan implementasi pada penelitian yang dilakukan oleh penulis mencakup kebutuhan perangkat keras dan perangkat lunak yang akan dijelaskan pada tabel 5.1 dan tabel 5.2 berikut ini :

**Tabel 5.1. Perangkat Keras**

No	Nama Perangkat	Spesifikasi
1	<i>Harddisk</i>	<i>1000 GB</i>
2	<i>Memory</i>	<i>8 GB</i>
3	<i>Processor</i>	<i>Core i7 2.80 GHz</i>
4	<i>Storage</i>	<i>SSD 250 GB</i>
5	<i>Monitor</i>	<i>15 inch</i>

**Tabel 5.2. Perangkat Lunak**

No	Nama Perangkat	Spesifikasi
1	<i>Development Tools</i>	<i>Google Colab dan Visual Studio</i>
2	<i>Programming Language</i>	<i>Python 3.7</i>
3	<i>Browser</i>	<i>Google Chrome</i>
4	<i>Documentation</i>	<i>Microsoft Word</i>

Pada tabel 5.2. dapat dilihat bahwa penulis menggunakan platform untuk memproses deep learning yaitu *Google Colab. Collaboratory*, atau disingkat "*Colab*", adalah produk *Google Research*. *Colab* memungkinkan siapa saja untuk menulis dan mengeksekusi kode python melalui browser secara online, dan sangat cocok untuk pembelajaran mesin, analisis data, dan pendidikan. Adapun hardware accelerator yang dipakai ialah *GPU*. Dapat dilihat pada gambar 5.1. Menggunakan *GPU* dapat menambah kecepatan proses pembelajaran mesin yang dilakukan dengan menambah RAM menjadi 15GB serta mendapatkan konsumsi *GPU* secara gratis yang sudah tersedia oleh *google colab*.



NVIDIA-SMI 470.42.01 Driver Version: 460.32.03 CUDA Version: 11.2									
GPU	Name	Persistence-M		Bus-Id	Disp.A	Volatile Uncorr. ECC			
Fan	Temp	Perf	Pwr:Usage/Cap	Memory-Usage		GPU-Util	Compute M.	MIG M.	
0	Tesla T4		Off	00000000:00:04.0	Off			0	
N/A	43C	P8	9W / 70W	0MiB / 15109MiB		0%	Default	N/A	

Processes:									
GPU	GI	CI	PID	Type	Process name	GPU Memory			
	ID	ID				Usage			
No running processes found									

Gambar 5.1. Summary GPU Google Colab

## 5.2. Pengujian

Pada tahap ini akan menjawab semua hal yang ada pada bab sebelumnya tentang metodologi penelitian. Dari mulai hasil proses scraping, pembersihan data, analisis sentimen sampai dengan evaluasi.

### 5.2.1. Hasil Scraping Data

Data diambil dari twitter menggunakan bahasa pemrograman Python, dengan *library* *snsrape* dan *tweepy* untuk terhubung ke Twitter *API*. Sebanyak kurang lebih 34000 tweet dipilih dengan kata kunci pencarian vaksin dan di ikuti dengan nama kota dari masing-masing provinsi. Pengambilan data twitter juga dikasih interval waktu dari awal bulan januari 2021 sampai akhir bulan juli 2021. Contoh hasil scraping data Twitter terlihat pada tabel 5.3 dibawah ini :

Tabel 5.3. Hasil Scraping Data Twitter

No	Username	Date	Content	User_location	Bulan
1	bellzia	2021-02-02 11:16:55+00:00	Hari ini vaksin dosis kedua (booster), gak kerasa ngantuk seperti dosis pertama kemarin, dan gak pegel 😊	Nusa Tenggara Barat	Februari
2	Andreebagaas	2021-01-14 17:13:42+00:00	Aku ndamau di suntik vaksin, vaksin ku adlh dzikir	Jawa Tengah	Januari
3	whitehorsegroup	2021-03-29 11:09:25+00:00	Seluruh driver & crew White Horse Bali sudah mendapat vaksin Covid-19.	DKI Jakarta	Maret

4	niczous	2021-04-28 13:52:36+00:00	Peluang untuk dapat vaksin awal! 🤔 <a href="https://t.co/HyXQwDfk08">https://t.co/HyXQwDfk08</a>	Aceh	April
5	dhido_blenk	2021-07-06 20:52:23+00:00	Buat teman-teman yang hendak berwisata ke Bali, jadi makin yakin kan buat naik Bus White Horse? 😊🙏 <a href="https://t.co/N1kZ6bl5de">https://t.co/N1kZ6bl5de</a>	Nusa Tenggara Timur	Januari

### 5.2.2. Labelling Data Manual

Pemberian label disini ialah untuk mendapatkan data *training* untuk melakukan pengujian akurasi pada metode yang akan dilakukan. Disini akan melakukan labelling data sebanyak 1000 data awal, dengan menggunakan pemahaman dari penulis maka didapatkan untuk label sentimen positif, negatif dan netral.

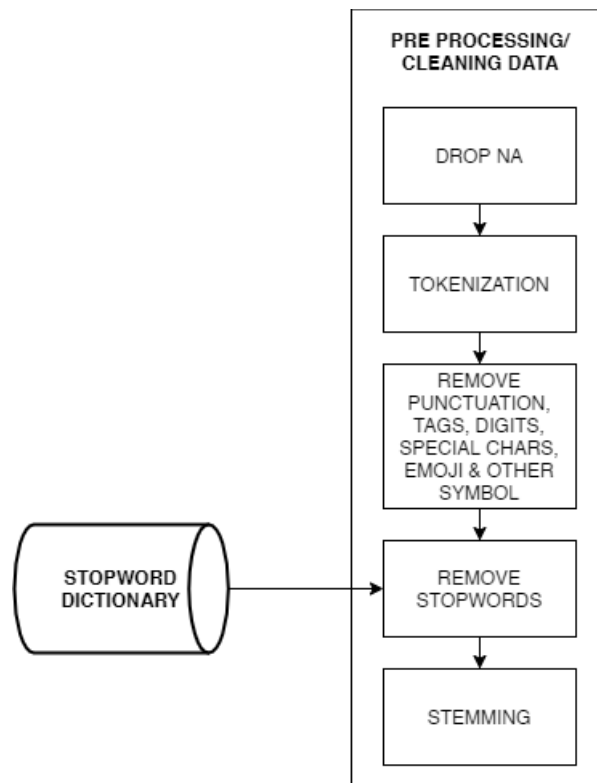
+-----+	+-----+	+-----+	
positive	negative	neutral	
+-----+	+-----+	+-----+	
570	317	172	
+-----+	+-----+	+-----+	

Gambar 5.2. Labelling Data Manual

Pada gambar 5.2 bahwa tweet yang diberi label secara manual ialah sebanyak 1059 data. Dengan mendapatkan hasil positif sebanyak 570, negatif 317 serta sentimen netral sebanyak 172.

### 5.2.3. Pre-procesing Data

Data yang telah didapatkan dari twitter umumnya berupa kalimat yang tidak terstruktur. Seperti contoh pada tabel 5.3, seperti karakter non-alfanumerik, *link url*, tata bahasa yang tidak beraturan, dll. Kemudian itu membuat adanya *noise* dan ketidakteraturan data yang dapat mempengaruhi kinerja proses *modelling* menggunakan Python nantinya.



Gambar 5.3. Langkah-Langkah *Pre-processing Data*

#### 5.2.2.1. *DROP NA*

Proses membersihkan data yang pertama ialah *DROP NA*. Dengan melakukan drop NA maka data akan menjadi lebih bersih karena jika ada data yang null atau kosong maka akan dihapus secara otomatis.

#### 5.2.2.2. *Tokenization*

Pada tahap ini melakukan proses penguraian teks menjadi perkata. Hasil ini disebut dengan token. Contohnya ada pada tabel 5.4 dibawah ini. Namun, sebelum melakukan *tokenization* penulis melakukan case folding untuk mengkonversi huruf kapital yang ada di data menjadi huruf kecil.

Tabel 5.4. Proses *Tokenization*

Teks Asli	<i>Tokenization</i>
ayo jangan takut vaksin	'ayo', 'jangan', 'takut', 'vaksin'

#### 5.2.2.3. *Remove Punctuation, Tags, Digits, Special Chars, Emoji & Other Symbol*

Dataset yang sudah berlabel kemudian dibersihkan dari karakter yang tidak relevan seperti tautan url, tagar, nama pengguna, tanda baca, dan karakter non-alfanumerik (kecuali spasi)

seperti ! @ # \$ & % dll. Karakter ini sangat mempengaruhi keakuratan proses analisis, karena mereka tidak menyertakan teks yang dapat ditemukan di kamus. Penghapusan karakter menggunakan pustaka ekspresi reguler python.

**Tabel 5.5. Penghapusan Karakter**

<b>Teks Asli</b>	<b>Setelah dihapus karakter yang tidak penting</b>
Seluruh driver & crew White Horse Bali sudah mendapat vaksin Covid-19.	'Seluruh driver amp crew White Horse Bali sudah mendapat vaksin Covid'

#### **5.2.2.4. Remove Stopwords**

Penerapan *stopwords* bertujuan untuk menghilangkan kata-kata yang dianggap tidak penting, yang dapat mempengaruhi kecepatan dan kinerja proses analisis (Anastasia & Budi, 2017). Proses stopwords dalam penelitian ini menggunakan library python yaitu Sastrawi. Beberapa kata pada stoplist python sastra adalah 'yang', 'untuk', 'di', 'dan', 'ke', 'dari', dll. Penulis menambahkan beberapa kata yang harus dihapus pada dataset yang penulis miliki seperti 'rt', 'rts', 'vaksin', 'covid', 'corona', 'korona', 'https', 'http', 'url'. Dengan menghapus kata-kata yang tidak penting tersebut data akan menjadi lebih bersih.

#### **5.2.2.5. Stemming**

*Stemming* merupakan cara untuk menghilangkan kata yang berupa awalan, akhiran atau campuran: (awalan-akhiran) menjadi bentuk kata dasar menurut kaidah bahasa Indonesia (Rohman et al., 2018). Proses *stemming* ini menggunakan library python *Sastrawi*. Sebuah library sederhana yang dirancang untuk mendapatkan kata-kata dasar bahasa Indonesia dengan mudah. Dalam tabel 5.6 Ada beberapa contoh proses stemming pada kata berimbuhan.

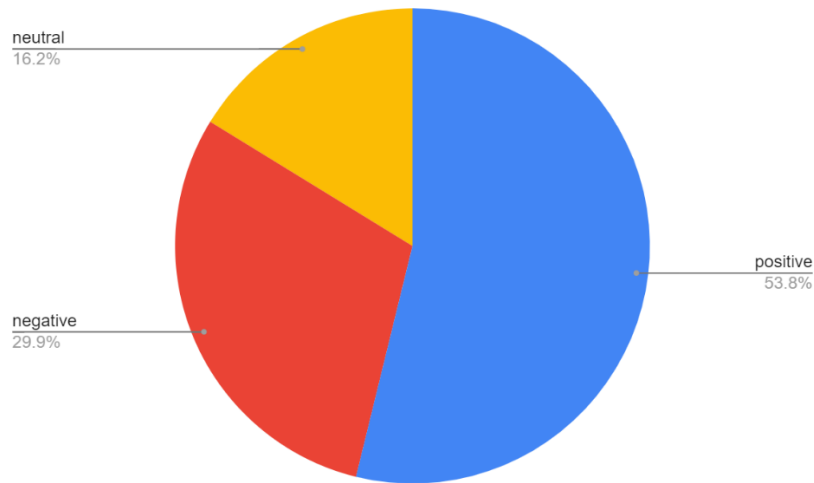
**Tabel 5.6. Proses Stemming**

<b>Sebelum di Stemming</b>	<b>Sesudah di Stemming</b>
Berkerumun Kerumunan	Kerumun
Menolak Penolakan	Tolak

#### **5.2.4. Sentiment Analysis**

Pada tahap ini dijelaskan seluruh pengolahan data untuk analisis sentimen dengan menggunakan metode *lexicon based* dan *CNN* dengan *word embedding* dari model *fasttext*. Hasil juga akan menghasilkan suatu akurasi dan evaluasi untuk mengukur kinerja yang penulis lakukan pada penelitian kali ini. Adapun hasil pengujiannya yaitu :

1. Pertama 1000 data akan di label secara manual dengan 3 kelas yaitu positif, negatif atau netral. Hasil ini akan menjadi perbandingan akurasi untuk metode *lexicon based*. Hasilnya dapat dilihat pada diagram *pie chart* berikut ini :



**Gambar 5.4. Pie Chart Labelling Manual**

2. Dilakukan *modelling lexicon based* dengan menggunakan 1000 data yang sudah di labelin secara manual. Dengan seperti ini maka didapat akurasi dengan membandingkan metode *lexicon based* dengan hasil yang secara manual. Metode *lexicon based* atau berbasis leksikon menggunakan kamus stok kata dengan kata-kata opini dan mencocokkan rangkaian kata yang diberikan dalam sebuah teks untuk menemukan polaritas. Contoh dari kamus *lexicon* bisa dilihat dari tabel 5.7 berikut :

**Tabel 5.7. Contoh Kamus *Lexicon Based***

teks	nilai polaritas
membantu	5
benar	4
tolong	3
tolol	-5
liang	-3
biadap	-5

Adapun hasil dari proses metode *lexicon based* sebagai berikut :

text	sentiment
ahyanuar tirtoid jokowi pakai indon maksud mal...	-1
tirtoid jokowi maksud sih turki batas masyarak...	0
konsep dosis dosis data efikasi johnson amp jo...	4
dirgarambe dok	4
enam jam pasca selesai praktek rapat kipi dink...	2

Gambar 5.5. Contoh Hasil *Lexicon Based*

Dari hasil *lexicon based* dapat dilihat bahwa sentimen masih mengandung nilai polaritas dari -15 sampai 15. Oleh karena itu, untuk membuat nilainya mengandung positif, negatif atau netral. Dilakukan proses pengubahan nilai menjadi sentimen sebagai berikut :

```
df['sentiment'].apply(lambda score: 'positive' if score>=0.01
else 'negative' if score<=-0.01 else 'neutral')
```

Dengan melakukan rumus di atas maka hasil akhir secara keseluruhan untuk melakukan prediksi sentimen menggunakan *lexicon based* untuk 1000 data di awal ialah sebagai berikut :

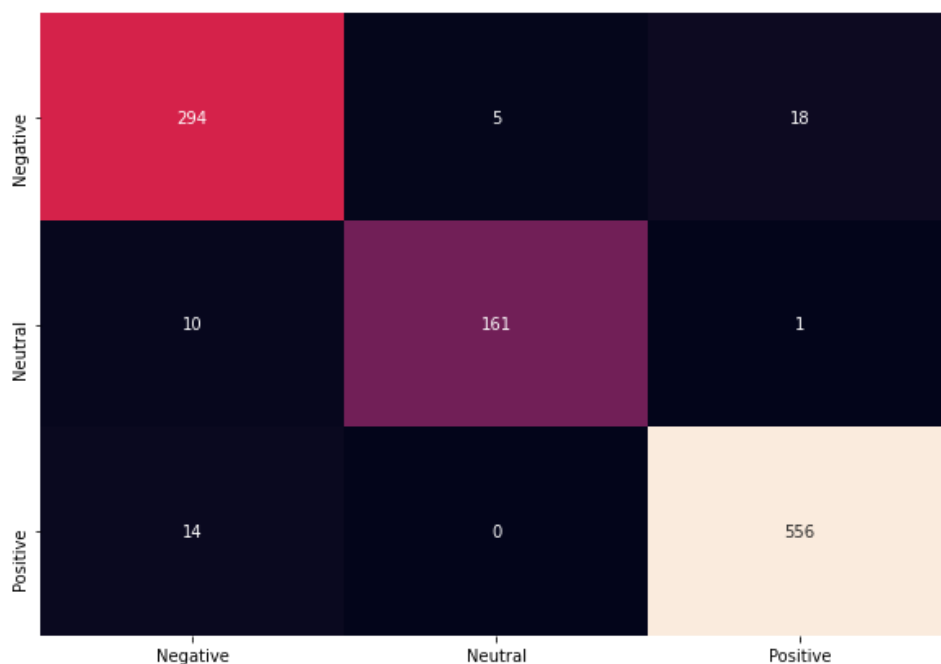
ahyanuar tirtoid jokowi pakai indon maksud mal...	-1	negative
tirtoid jokowi maksud sih turki batas masyarak...	0	neutral
konsep dosis dosis data efikasi johnson amp jo...	4	positive
dirgarambe dok	4	positive
enam jam pasca selesai praktek rapat kipi dink...	2	positive

Gambar 5.6. Contoh Hasil *Lexicon Based* Setelah di sentimen

Adapun akurasi untuk tahap pertama menggunakan *lexicon based* adalah 95% dengan membandingkan labelling manual dengan hasil labelling menggunakan *lexicon based*. Hasil dari akurasi dapat dilihat pada gambar 5.7 yaitu classification report dan 5.8 terkait confusion matrix.

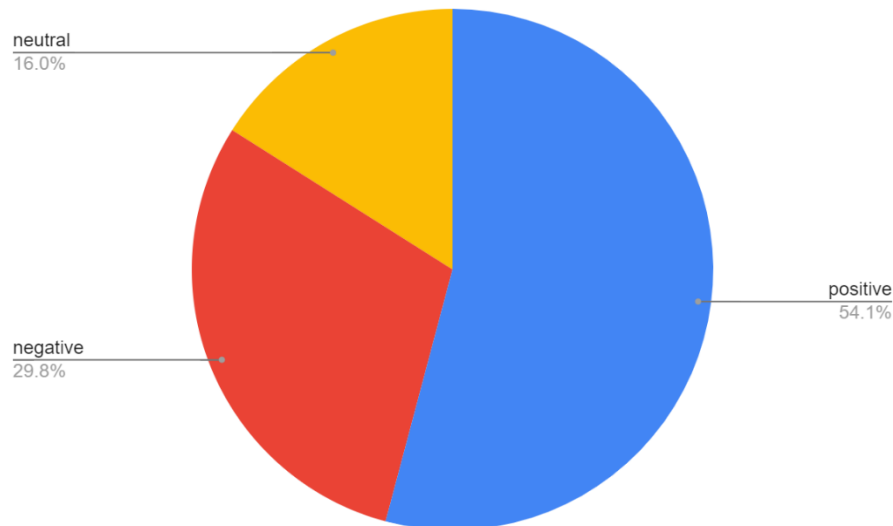
	precision	recall	f1-score	support
negative	0.92	0.93	0.93	317
neutral	0.97	0.94	0.95	172
positive	0.97	0.98	0.97	570
accuracy			0.95	1059
macro avg	0.95	0.95	0.95	1059
weighted avg	0.95	0.95	0.95	1059

**Gambar 5.7. Classification Report Lexicon Based**



**Gambar 5.8. Confussion Matrix Lexicon Based**

- Hasil dari tahap kedua jika didapatkan akurasi sebesar 95% maka sudah termasuk kedalam akurasi yang cukup bagus untuk melakukan label otomatis dengan 5000 data.
- Dilakukan *modelling lexicon based* dengan 5000 data.
- Kemudian pada tahap ini, 1000 data yang dilakukan label secara manual di *merge* atau digabungkan dengan 5000 data yang sudah diberi label secara otomatis menggunakan *lexicon based*. Hasilnya didapat 6000 data untuk dijadikan *dataset training CNN*. Untuk mengetahui persebaran sentimennya dapat dilihat pada diagram *pie chart* pada gambar berikut ini :



**Gambar 5.9. Pie Chart Hasil Lexicon Based**

Pada gambar 5.2 bahwa tweet yang diberi label menggunakan metode *lexicon based* ialah sebanyak 6037 data. Dengan mendapatkan hasil positif sebanyak 3268, negatif 1802 sentimen netral sebanyak 967. Maka, dengan seperti ini 6000 data yang sudah diberi label akan menjadi dataset *training CNN*.

6. Melakukan *modelling CNN* dengan beberapa Teknik yang ada didalamnya yaitu, dengan melakukan perbandingan beberapa model dalam *word embedding* yang penulis gunakan *fasttext*, *word2vec* dan *GloVe*. Pengujian ini menggunakan 1000 data diawal. Hasil dari perbandingan tersebut dapat dilihat pada tabel 5.8 berikut ini :

**Tabel 5.8. Perbandingan Word Embedding**

<i>model</i>	<i>Accuracy</i>	<i>Epoch</i>
<i>CNN Without Pre-Trained Word Embedding</i>	87%	100
<i>CNN + Fasttext with cbow</i>	93%	100
<b><i>CNN + Fasttext with skipgram</i></b>	<b>98%</b>	<b>100</b>
<i>CNN + Word2vec</i>	87%	100
<i>CNN + GloVe</i>	88%	100

Berdasarkan hasil evaluasi kinerja ketiga *word embedding*, *Fasttext* lebih unggul dibandingkan *word2vec* dan *glove*. *Word2vec* dan *GloVe* adalah pre-trained embedding yang mendefinisikan sebuah fungsi yang memetakan sebuah kosakata  $v$  ke dalam vektor  $h$ .

$$f_{vocab} : v \rightarrow h \quad (1)$$

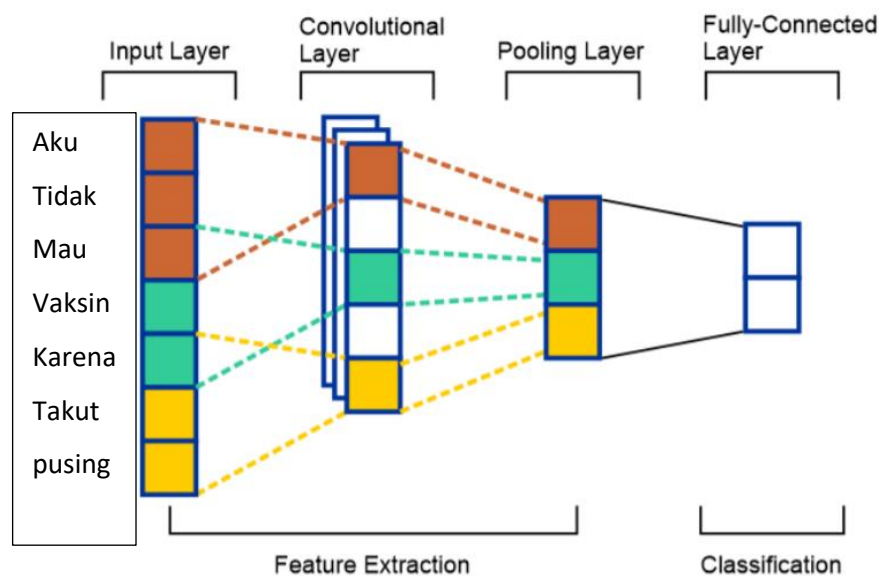


Sedangkan *FastText* memetakan suku kata dari sebuah kosakata dan urutan karakter ( $c_1 \dots c_n$ ) ke dalam vektor  $h$ . Urutan karakter dari bahasa seringkali mengindikasikan komposisi informasi dari makna sebuah kata.

$$f_{\text{subword}} : (v(c_1 \dots c_n)) \rightarrow h \quad (2)$$

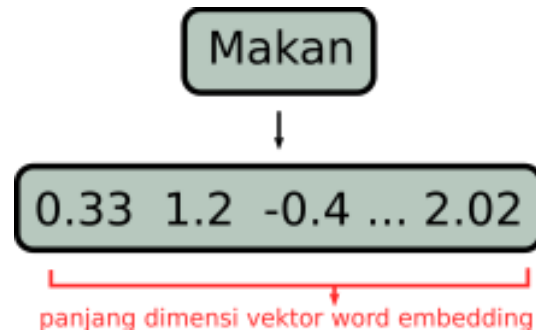
*FastText* memiliki kinerja yang baik, dapat melatih model pada dataset yang besar dengan cepat dan dapat memberikan representasi kata yang tidak muncul dalam data latih. Jika kata tidak muncul selama pelatihan model, kata tersebut dapat dipecah menjadi n-gram untuk mendapatkan *embedding* vektornya.

- Setelah mendapatkan akurasi mana yang lebih tinggi untuk word embedding maka selanjutnya adalah dengan membuat arsitektur dari metode *CNN*. Metode yang dipakai ialah *Convolutional Neural Network* dengan word embedding *Fasttext*. *CNN* merupakan bagian dari *Artificial Neural Network (ANN)* yang mampu mendeteksi informasi dengan akurasi tinggi (Rhanoui et al., 2019). Model *CNN* telah memecahkan permasalahan dalam pemrosesan gambar dan saat ini peneliti-peneliti telah mengembangkan *CNN* untuk *NLP* seperti analisis sentimen, klasifikasi polaritas emosional, *text summary*, dll (Luo, 2019).



Gambar 5.10. Contoh Arsitektur *CNN*

8. Ekstraksi Fitur di penelitian ini menggunakan model *word embedding* dari *fasttext*. Fasttext disini berperan dalam mengubah teks korpus menjadi vektor dengan mengubah setiap teks menjadi susunan urutan integer, dimana setiap integer akan menjadi indeks dari kamus token. Model *word embedding* dari fasttext ini menggunakan *library gensim*.



Gambar 5.11. Contoh Word Embedding Fasttext

9. Dengan menggunakan 6000 data yang telah diberi label maka akan dilakukan *training* data menggunakan *CNN*. Sebelum melakukan *training data*, akan dilakukan *splitting* data. Dalam menguji sebuah model diperlukan pembagian dataset yaitu data latih dan data uji. Untuk melakukan pembagian dataset digunakan the *SciKit library* dengan nama kelas '*train\_test\_split*'. Dengan menggunakan '*train\_test\_split*' dapat membagi data latih dan data uji secara acak menjadi dalam berbagai proporsi.

```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x, y,
test_size=0.20, random_state=42)
```

Pada kode diatas maka data dibagi menjadi data train 80% dan data test 20%

10. Membangun Model *CNN*. Tipe model yang digunakan pada kasus ini yaitu *Sequential*. Dengan *Sequential*, membangun model dapat dibuat dari layer ke layer secara berurutan. Setiap layernya memiliki bobot yang bersesuaian dengan layer berikutnya. Penambahan layer pada model dapat ditambahkan dengan mudah menggunakan fungsi '*add()*'. '*Activation*' adalah fungsi aktivasi untuk lapisan model. Tujuan dari aktivasi yaitu untuk mengatasi kasus rumit dan non linier yang kompleks. Fungsi aktivasi yang digunakan yaitu *ReLU* atau *Rectified Linear Activation*, *Sigmoid* dan *Softmax*.

Pada model ini digunakan *ReLU* karena merupakan satu dari beberapa fungsi aktivasi yang terbaik (Wallace, 2014). Layer terakhir yang digunakan yaitu *layer Dense* atau *Fully Connected*

layer sebagai lapisan keluaran. Dalam *Dense layer*, semua node di lapisan sebelumnya terhubung ke *node* di lapisan saat ini. Pada layer ini, untuk model sentimen memanfaatkan fungsi aktivasi *sigmoid* karena fungsi ini yang paling sesuai untuk kategori 3 kelas. Untuk *compile model* dibutuhkan 2 parameter, yaitu *optimizer* dan *loss*. *Optimizer* bertugas untuk mengontrol *learning rate*. Pada kasus ini, digunakan *optimizer* ‘adam’ yang dianggap *optimizer* bagus dan sering digunakan. Untuk fungsi *loss*, digunakan *binary\_crossentropy* pada model sentimen karena paling sesuai untuk kategori dengan 3 kelas. Pada gambar 5.12 dan 5.13 merupakan code program dan *summary* dari model yang akan di bangun.

```
model = Sequential()
model.add(Embedding(nb_words+1, embed_dim,
                    weights=[embedding_matrix],
                    input_length=max_seq_len,
                    trainable=False))
model.add(Conv1D(num_filters, 7, activation='relu', padding='same'))
model.add(MaxPooling1D(2))
model.add(BatchNormalization())
model.add(Conv1D(num_filters, 7, activation='relu', padding='same'))
model.add(GlobalMaxPooling1D())
model.add(BatchNormalization())
model.add(Dropout(0.5))
model.add(Dense(32, activation='relu', kernel_regularizer=regularizers.l2(weight_decay)))
model.add(Dense(num_classes, activation='softmax'))

adam = optimizers.Adam(lr = 0.01)

model.compile(loss='categorical_crossentropy', optimizer='adam', metrics=['accuracy'])
model.summary()
```

**Gambar 5.12. Kode Program Model CNN**

```

=====
>>> Training CNN >>>
=====
Model: "sequential_2"

```

Layer (type)	Output Shape	Param #
embedding_2 (Embedding)	(None, 55, 100)	1613200
conv1d_2 (Conv1D)	(None, 55, 64)	44864
max_pooling1d_1 (MaxPooling1D)	(None, 27, 64)	0
batch_normalization_2 (Batch Normalization)	(None, 27, 64)	256
conv1d_3 (Conv1D)	(None, 27, 64)	28736
global_max_pooling1d_1 (GlobalMaxPooling1D)	(None, 64)	0
batch_normalization_3 (Batch Normalization)	(None, 64)	256
dropout_1 (Dropout)	(None, 64)	0
dense_2 (Dense)	(None, 32)	2080
dense_3 (Dense)	(None, 3)	99

```

=====
Total params: 1,689,491
Trainable params: 76,035
Non-trainable params: 1,613,456

```

**Gambar 5.13. Summary Program Model CNN**

## 11. Training Model CNN

Pelatihan model *deep learning* menggunakan fungsi `'fit()'` yang diikuti 5 parameter: training data (`train_X`), target data (`train_y`), *validation split*, dan *callbacks*. Penggunaan *validation split* akan membagi data secara acak menjadi data latih dan data uji. Pada kasus ini nilai *validation split* yaitu 0.20, sehingga 20% data dari data *training* akan digunakan untuk menguji performa model. Ketika pelatihan data, dengan adanya pembagian data akan muncul *validation loss*. *Validation loss* yaitu *error* yang muncul setelah menjalankan data *validation* melalui model.

```

hist = model.fit(word_seq_train,
                  train_labels,
                  batch_size = batch_size,
                  epochs = num_epochs,
                  validation_split=0.2,
                  callbacks = callbacks_list,
                  shuffle=True,
                  verbose=2)

```

**Gambar 5.14. Kode Program Training Model CNN**

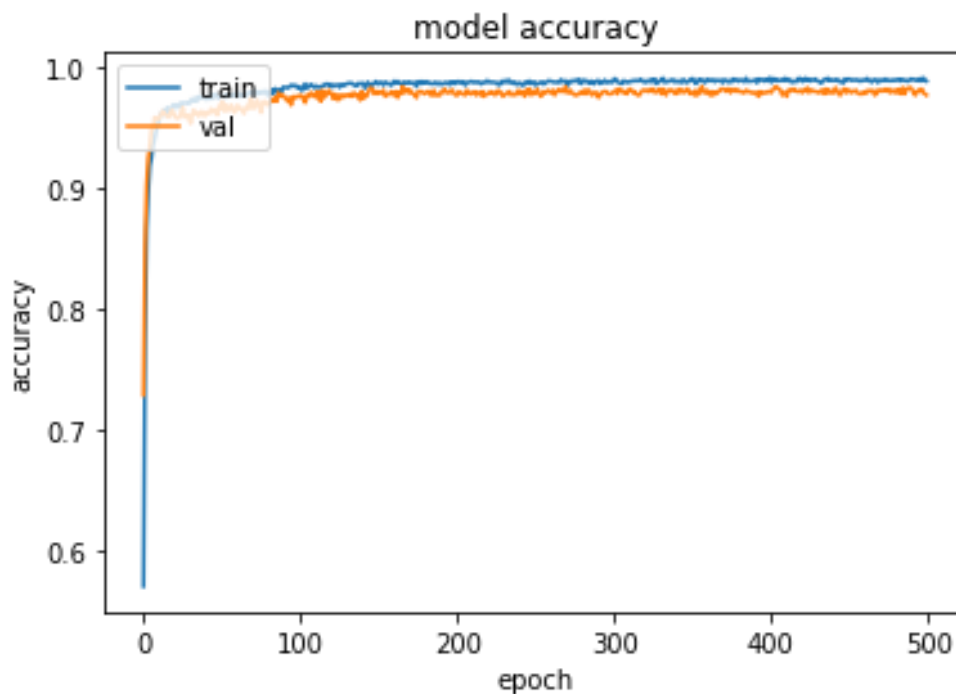
*Epoch* adalah berapa kali model akan melakukan *training* data. Di penelitian ini akan diulang melakukan *training* dengan 6000 data. Kemudian dilakukan perbandingan *epoch* 10, 20, 50, 100 dan 500 epoch. Hasilnya adalah sebagai berikut:

Epoch	Accuracy
10	79 %
20	80 %
50	77 %
100	89 %
500	97 %

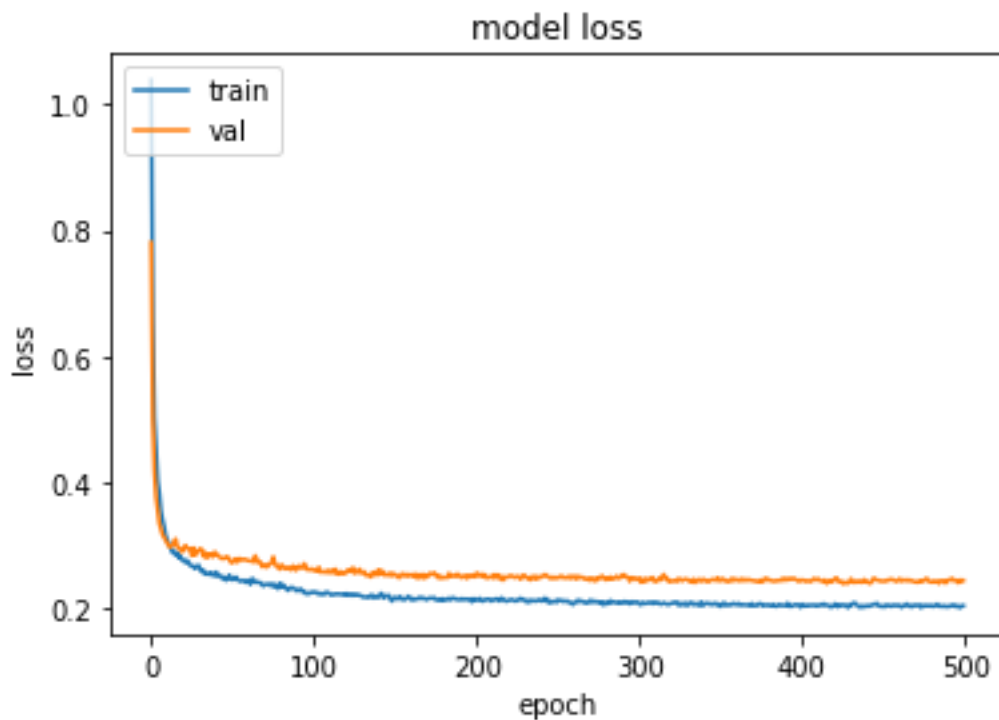
**Gambar 5.15. Perbandingan *Epoch Model Training CNN***

Dari gambar 5.15 dapat diketahui bahwa di *epoch* 500 menghasilkan akurasi sebesar 97% dan yang dipahami penulis itu sudah cukup untuk menghasilkan suatu prediksi terhadap data testing sebanyak 28000 data lagi.

## 12. Evaluasi Model



**Gambar 5.16. Akurasi *Model CNN***



**Gambar 5.17. Loss Function Model CNN**

Pada gambar 5.16 dilihat akurasi dari *epoch* 0 sampai 500 mengalami peningkatan secara terus-menerus artinya disini model memiliki akurasi yang baik. Pada gambar 5.17 dilihat bahwa *loss* mengalami penurunan, artinya disini model memiliki jumlah *loss* yang sedikit yang berarti model memiliki akurasi yang baik antara *loss* dan juga akurasi.

### 5.3. Analisis Hasil Pengujian

#### 5.3.1. Prediksi Data Baru

Sebelumnya model yang telah ditraining di *save* dengan tipe *.h5* dan untuk melakukan prediksi data baru yang penulis lakukan kepada 22270 data.

```
model = load_model('/content/drive/MyDrive/ngoding/skripsi/hasil/model_
fasttext2/new_model2.h5')
```

Pada code diatas melakukan *load model* yang telah di *save* sebelumnya. Kemudian, melakukan *load* untuk *word embedding* yang sebelumnya juga sudah di *save* dengan tipe *.pkl*

```
with open('/content/drive/MyDrive/ngoding/skripsi/hasil/model_fasttext2
/new_tokenizer.pickle', 'rb') as handle:
    tokenizer = pickle.load(handle)
```

Dengan seperti itu, maka penulis melakukan prediksi dengan memanggil data csv.

```
y_pred = model.predict(word_seq_test).round()
```

Pada kode diatas maka kita melakukan prediksi dengan menggunakan model yang sudah di load sebelumnya. Maka mendapatkan gambaran hasil prediksi sebagai berikut :

	text	predict
0	*POLRES BARITO SELATAN* \n\nDalam rangka mendu...	1
1	Warga Banjarmasin, Kalimantan Selatan (Kalsel)...	1
2	Polresta Banjarmasin vaksin 172 jiwa termasuk ...	1
3	DIR LANTAS POLDA KALSEL KOMBES POL MAESA SOEGR...	1
4	@jokowi Pak Perseden saya dari kalimantan sela...	1
...	...	...
22265	Jaga Kesehatan Diri Dengan Vaksin COVID-19\n\n...	1
22266	Polres Jayapura, Papua membongkar sindikat pel...	0
22267	POLRES JAYAPURA UNGKAP PELAKU PEMALSUAN SURAT ...	0
22268	Polisi menyita barang bukti berupa laptop, pri...	1
22269	RSUD 2 JAYAPURA\n..SIAPA YANG Bertanggung jawab...	2

22270 rows × 2 columns

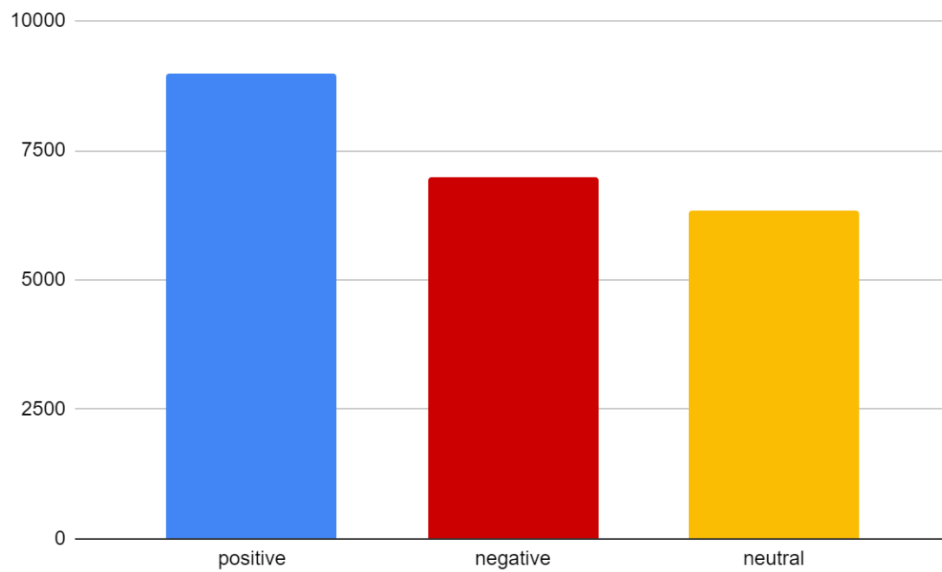
**Gambar 5.18. Prediksi Data Baru**

```
1 predict['predict'].value_counts()

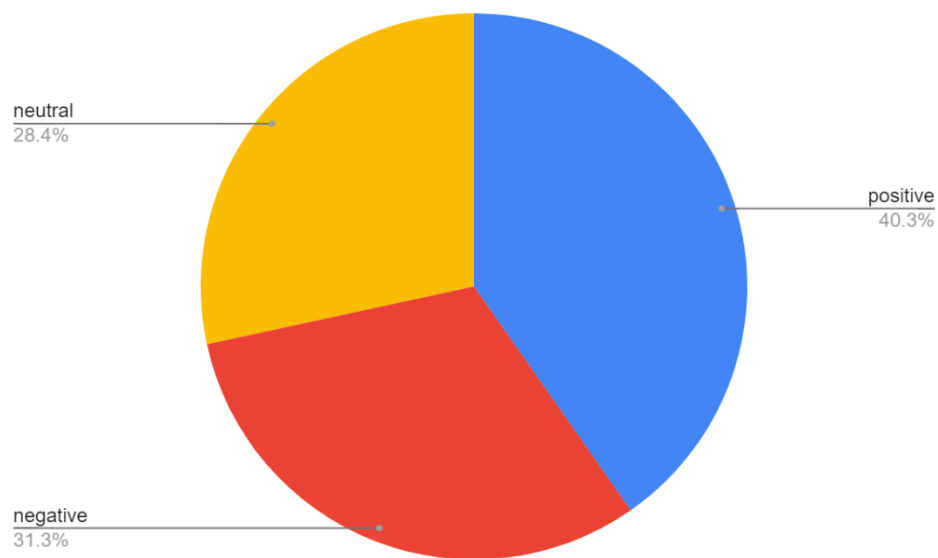
1      8971
2      6974
0      6325
Name: predict, dtype: int64
```

**Gambar 5.19. Hasil Sentimen Prediksi Data Baru**

Pada gambar diatas maka dapat disimpulkan ada angka 1 yaitu positif sebanyak 8971, angka 2 yaitu negatif sebanyak 6974 dan angka 0 yaitu netral sebanyak 6325. Dengan seperti ini maka hasil analisis sentimen dari bulan januari – juli 2021 mendapatkan hasil analisis sentimen lebih tinggi positif.



**Gambar 5.20. Perbandingan Hasil Sentimen Prediksi**

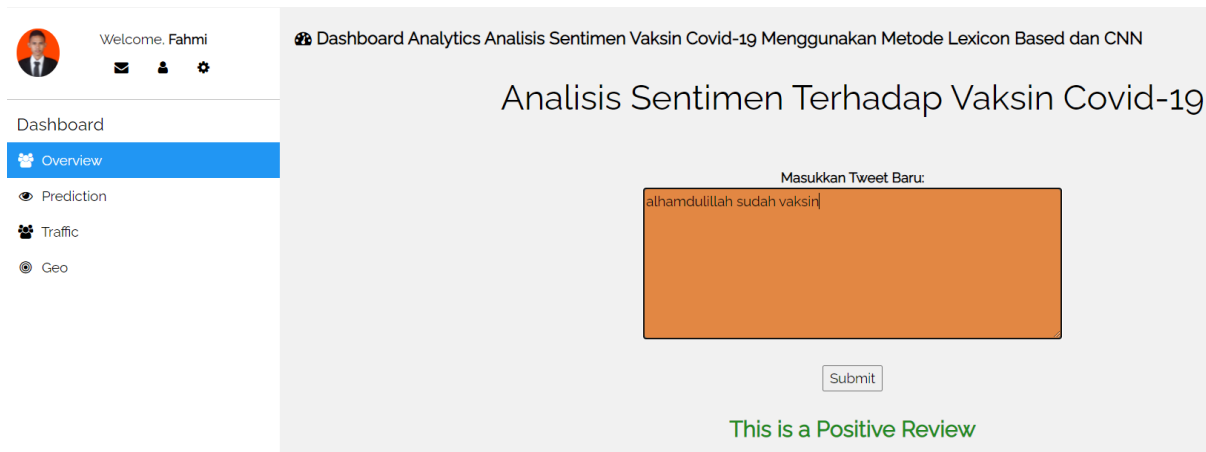


**Gambar 5.21. Diagram Pie Chart Hasil Sentimen Prediksi**

### 5.3.2. Deployment Model Menggunakan Framework Flask

Untuk melakukan deployment model penulis menggunakan *framework flask* dari python yang sudah banyak digunakan pada peneliti untuk melakukan *deployment model* machine learning ataupun *deep learning*. Berikut untuk input teks dan output sentimen.

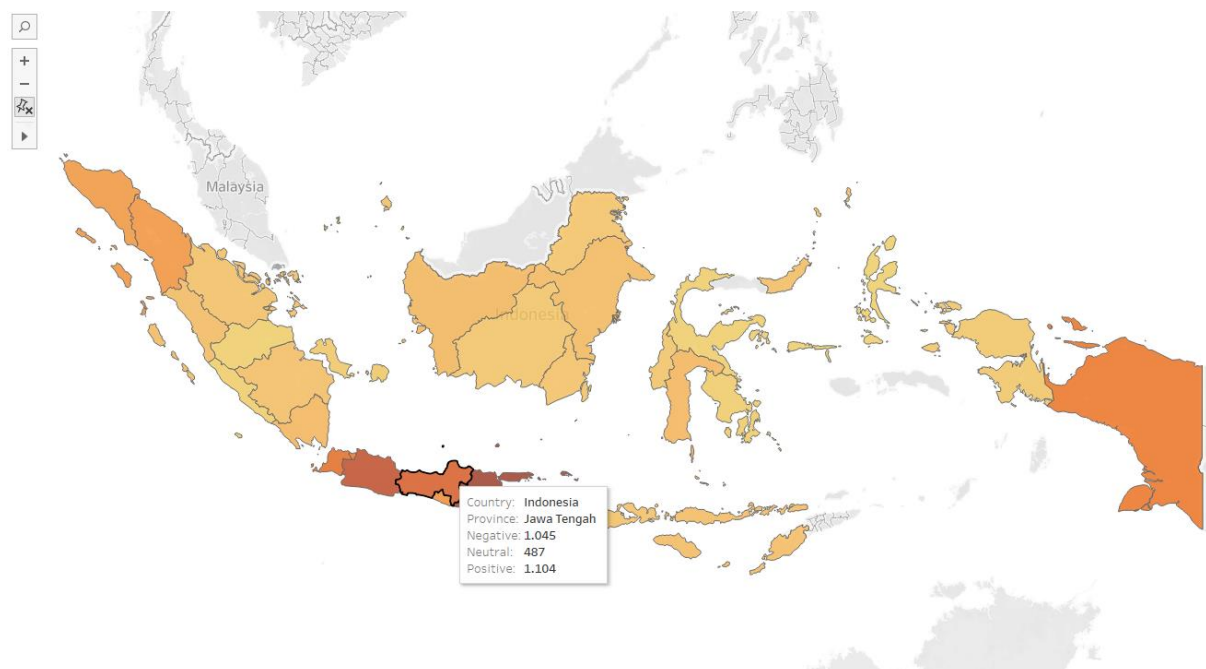




Gambar 5.22. Tampilan Halaman Prediksi Teks Baru

### 5.3.3. Visualisasi Menggunakan *Tableau*

Ada salah satu kelebihan dari *tableau* untuk membuat visualisasi yaitu dapat membuat persebaran data berbentuk *maps* dengan mengambil *longitude* dan *latitude* dari setiap provinsi. Ini menjadi hal yang membuat penelitian ini menjadi lebih mudah untuk dibaca analisisnya. Adapun tampilan visualisasi mapsnya ialah seperti ini:



Gambar 5.23. Tampilan Maps Persebaran Data

### 5.3.3. *Tableau To Web*

Setelah melakukan *deployment* dan beberapa visualiasi menggunakan *tableau* maka Langkah terakhir yaitu membuat visualisasi itu bisa tampil di halaman *web* yang di *deploy* menggunakan *framework flask* dari *python*.

## **BAB VI**

### **KESIMPULAN DAN SARAN**

Ada beberapa metode yang dapat digunakan untuk melakukan analisis sentimen salah satunya ialah yang digunakan penulis ialah Teknik dari deep learning yaitu *Convolutional Neural Network* atau yang biasa disingkat dengan *CNN*. Untuk membuat akurasi *CNN* menjadi lebih tinggi diperlukan word embedding untuk merepresentasikan teks ke dalam bentuk vektor, salah satunya dengan menggunakan *word embedding*. Di penelitian ini penulis membandingkan *CNN* dengan beberapa metode *word embedding* yaitu *fasttext*, *word2vec* dan *glove*. Adapun kesimpulan dan saran dari penelitian ini ialah :

#### **6.1 Kesimpulan**

1. Nilai akurasi pada model *CNN – Fasttext* lebih tinggi dibandingkan model yang lainnya yang memakai *word2vec* dan *glove*. Akurasi yang didapat ialah sebesar 97%.
2. Akurasi yang didapat untuk model *CNN-word2vec* ialah 87% dan *CNN-GloVe* ialah sebesar 88%.
3. Nilai Akurasi untuk pembuatan model *word embedding* lebih tinggi model *Fasttext* dibandingkan *model embedding* lainnya.
4. Hasil analisis sentimen terhadap vaksin covid-19 pada penelitian ini menghasilkan sentimen positif sebanyak 8971, negative sebanyak 6974 dan netral sebanyak 6325. Dengan seperti ini maka sentimen pengguna twitter terhadap vaksin covid-19 dalam interval waktu januari – juli 2021 ialah bersentimen positif.

#### **6.2 Saran**

1. Pada penelitian selanjutnya dapat menambahkan jumlah dataset yang digunakan dan melakukan perbandingan akurasi berdasarkan jumlah dataset yang lebih banyak.
2. Pada penelitian selanjutnya dapat menambahkan source pengambilan data tidak hanya dari media sosial twitter agar mendapatkan data yang lebih bervariasi.
3. Pada penelitian selanjutnya dapat mengkombinasikan model-model algoritma neural network yang lain untuk mendapatkan hasil akurasi yang lebih baik.

## DAFTAR PUSTAKA

- Abd, D. H., Abbas, A. R., & Sadiq, A. T. (2021). Analyzing sentiment system to specify polarity by lexicon-based. *Bulletin of Electrical Engineering and Informatics*, 10(1), 283–289. <https://doi.org/10.11591/eei.v10i1.2471>
- Al-Sharif, E., Strianese, D., AlMadhi, N. H., D’Aponte, A., dell’Omo, R., Di Benedetto, R., & Costagliola, C. (2021). Ocular tropism of coronavirus (CoVs): a comparison of the interaction between the animal-to-human transmitted coronaviruses (SARS-CoV-1, SARS-CoV-2, MERS-CoV, CoV-229E, NL63, OC43, HKU1) and the eye. *International Ophthalmology*, 41(1), 349–362. <https://doi.org/10.1007/s10792-020-01575-2>
- Bagheri, H., & Islam, M. J. (2017). Sentiment analysis of twitter data. *ArXiv*, October. <https://doi.org/10.4018/ijhisi.2019040101>
- Cambria, E., Poria, S., Gelbukh, A., Nacional, I. P., & Thelwall, M. (2017). AFFECTIVE COMPUTING AND SENTIMENT ANALYSIS Sentiment Analysis Is a Big Suitcase. *Ieee Intelligent Systems*.
- Chadijah, S., Suyadi, A., & Tohadi, T. (2020). Tarik Menarik Kewenangan Pemerintah Pusat Dan Pemerintah Daerah Dalam Penanganan Pandemi Covid-19. ... : *Jurnal Ilmu Hukum*, 3(2), 226–236. <http://www.openjournal.unpam.ac.id/index.php/rjih/article/view/8091>
- Chi, G., & Cushman, M. (2020). Attention and citation: Common interests of researchers and journals. *Research and Practice in Thrombosis and Haemostasis*, 4(3), 353–356. <https://doi.org/10.1002/rth2.12322>
- Christina, S., & Ronaldo, D. (2020). Studi Literatur Sistematis Terhadap Pengembangan Leksikon Sentiment. *Jurnal ELTIKOM*, 4(2), 121–131. <https://doi.org/10.31961/eltikom.v4i2.211>
- Cormier, M., & Cushman, M. (2021). Innovation via social media – The importance of Twitter to science. *Research and Practice in Thrombosis and Haemostasis*, March. <https://doi.org/10.1002/rth2.12493>
- Cotfas, L. A., Delcea, C., Roxin, I., Ioanas, C., Gherai, D. S., & Tajariol, F. (2021). The Longest Month: Analyzing COVID-19 Vaccination Opinions Dynamics from Tweets in the Month following the First Vaccine Announcement. *IEEE Access*, 9, 33203–33223. <https://doi.org/10.1109/ACCESS.2021.3059821>
- D’Andrea, E., Ducange, P., Bechini, A., Renda, A., & Marcelloni, F. (2019). Monitoring the public opinion about the vaccination topic from tweets analysis. *Expert Systems with Applications*, 116, 209–226. <https://doi.org/10.1016/j.eswa.2018.09.009>
- Dongo, I., Cadinale, Y., Aguilera, A., Martínez, F., Quintero, Y., & Barrios, S. (2020). *Web Scraping versus Twitter API*. 263–273. <https://doi.org/10.1145/3428757.3429104>
- Emanuel, E. J., Luna, F., Schaefer, G. O., Tan, K. C., & Wolff, J. (2021). Enhancing the WHO’s proposed framework for distributing COVID-19 vaccines among countries. *American Journal of Public Health*, 111(3), 371–373. <https://doi.org/10.2105/AJPH.2020.306098>
- Gunnarsson, K., & Herber, O. (2020). *The Most Popular Programming Languages of GitHub’s Trending Repositories*.
- H. Manguri, K., N. Ramadhan, R., & R. Mohammed Amin, P. (2020). Twitter Sentiment

- Analysis on Worldwide COVID-19 Outbreaks. *Kurdistan Journal of Applied Research*, 54–65. <https://doi.org/10.24017/covid.8>
- Hamdan, H. (2016). *Université d ' Aix-Marseille Sentiment Analysis in Social Media*. May.
- Henry, D. (2021). *TwScraper: A Collaborative Project to Enhance Twitter Data Collection*. November 2017, 886–889. <https://doi.org/10.1145/3437963.3441716>
- Hussein, D. M. E. D. M. (2018). A survey on sentiment analysis challenges. *Journal of King Saud University - Engineering Sciences*, 30(4), 330–338. <https://doi.org/10.1016/j.jksues.2016.04.002>
- Ke, Q., Ahn, Y., & Sugimoto, C. R. (2016). A Systematic Identification of Scientists on Twitter 1. *STI 2016: Peripheries, Frontiers and Beyond*, September, 1–5.
- Kusumawati, I. 2017. (2017). Analisa Sentimen Menggunakan Lexicon Based Kenaikan Harga Rokok Pada Media Sosial Twitter. *Analisa Sentimen Menggunakan Lexicon Based Untuk Melihat Persepsi Masyarakat Terhadap Kenaikan Harga Rokok Pada Media Sosial Twitter*.
- Makris, M. (2020). Staying updated on COVID-19: Social media to amplify science in thrombosis and hemostasis. *Research and Practice in Thrombosis and Haemostasis*, 4(5), 722–726. <https://doi.org/10.1002/rth2.12410>
- Meena, R. (2020). *Russia ' s Covid – 19 Vaccine : Social discussion and rst emotions*. 1–13.
- Pokharel, B. P. (2020). Twitter Sentiment Analysis During Covid-19 Outbreak in Nepal. *SSRN Electronic Journal*, January. <https://doi.org/10.2139/ssrn.3624719>
- Pragholapati, A. (2020). *Covid-19 Impact on Students*. 1–6. <https://doi.org/10.35542/osf.io/895ed>
- Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza : A Python natural language processing toolkit for many human languages. *ArXiv*. <https://doi.org/10.18653/v1/2020.acl-demos.14>
- Rachman, F. F., & Pramana, S. (2020). Analisis Sentimen Pro dan Kontra Masyarakat Indonesia tentang Vaksin COVID-19 pada Media Sosial Twitter. *Health Information Management Journal*, 8(2), 100–109. <https://inohim.esaunggul.ac.id/index.php/INO/article/view/223/175>
- Rahayu, R. N. & S. (2021). Vaksin covid 19 di indonesia : analisis berita hoax. *Jurnal Ekonomi, Sosial & Humaniora*, 2(07), 39–49. <https://www.jurnalintelektiva.com/index.php/jurnal/article/view/422>
- Raschka, S., Patterson, J., & Nolet, C. (2020). Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence. *Information (Switzerland)*, 11(4). <https://doi.org/10.3390/info11040193>
- Sagala, H. G. (2020). Peran keluarga dan pasien dalam meningkatkan keselamatan dan pencegahan covid 19. *Journal Kesehatan*, 4(2), 1–8.
- Sallam, M. (2021). Covid-19 vaccine hesitancy worldwide: A concise systematic review of vaccine acceptance rates. *Vaccines*, 9(2), 1–15. <https://doi.org/10.3390/vaccines9020160>
- Sitepu, A., & Syafril, S. (2020). Case Report : A Confirmed COVID-19 in Newly Diagnosed Diabetes Mellitus Patient. *Journal of Endocrinology, Tropical Medicine, and Infectious Disease (JETROMI)*, 02(3), 144–152.
- Suryadi, D. (2021). Does it make you sad? A lexicon-based sentiment analysis on COVID-19

- news tweets. *IOP Conference Series: Materials Science and Engineering*, 1077(1), 012042. <https://doi.org/10.1088/1757-899x/1077/1/012042>
- Susilo, A., Rumende, C. M., Pitoyo, C. W., Santoso, W. D., Yulianti, M., Herikurniawan, H., Sinto, R., Singh, G., Nainggolan, L., Nelwan, E. J., Chen, L. K., Widhani, A., Wijaya, E., Wicaksana, B., Maksum, M., Annisa, F., Jasirwan, C. O. M., & Yuniastuti, E. (2020). Coronavirus Disease 2019: Tinjauan Literatur Terkini. *Jurnal Penyakit Dalam Indonesia*, 7(1), 45. <https://doi.org/10.7454/jpdi.v7i1.415>
- Taboada, M., Brooke, J., & Voll, K. (2021). *Lexicon-Based Methods for Sentiment Analysis. September 2010*.
- Trajkova, M., Alhakamy, A., Cafaro, F., Vedak, S., Mallappa, R., & Kankara, S. R. (2020). Exploring casual COVID-19 data visualizations on Twitter: Topics and challenges. *Informatics*, 7(3), 1–22. <https://doi.org/10.3390/INFORMATICS7030035>
- Vallat, R. (2018). Pingouin: statistics in Python. *Journal of Open Source Software*, 3(31), 1026. <https://doi.org/10.21105/joss.01026>
- Ward, J. W., & del Rio, C. (2020). The COVID-19 Pandemic: An Epidemiologic, Public Health, and Clinical Brief. *Clinical Liver Disease*, 15(5), 170–174. <https://doi.org/10.1002/cld.973>
- Wunderlich, F., & Memmert, D. (2020). Innovative approaches in sports science-Lexicon-based sentiment analysis as a tool to analyze sports-related twitter communication. *Applied Sciences (Switzerland)*, 10(2). <https://doi.org/10.3390/app10020431>
- Xu, J., Li, Y., Gan, F., Du, Y., & Yao, Y. (2020). Salivary Glands: Potential Reservoirs for COVID-19 Asymptomatic Infection. *Journal of Dental Research*, 99(8), 989. <https://doi.org/10.1177/0022034520918518>
- Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4). <https://doi.org/10.1002/widm.1253>

# LAMPIRAN

## Lampiran 1. Bukti Kartu Bimbingan (KAMBING)



POLITEKNIK POS INDONESIA  
PROGRAM STUDI D4 TEKNIK INFORMATIKA  
JL. SARIASIH NO. 54 BANDUNG 40151  
Telp. 022-2009562, 2009570  
Fax. 022-2009568

### FORMULIR KEGIATAN

#### TUGAS AKHIR

TA. 2020/2021

Nama : Muhammad Fahmi  
Npm : 1174021  
Judul : Analisis Sentimen Vaksin Covid-19 Berdasarkan Lokasi Di Indonesia Menggunakan Metode Lexicon Based Dan CNN  
Pembimbing : Rolly Maulana Awangga S.T., M.T.



Pertemuan	Tanggal	Sudah Dikerjakan	Pekerjaan Selanjutnya	Nilai
1	12-07-2021	judul dan metode ta	submit jurnal i1	100
2	22-07-2021	submit jurnal i1	lanjutkan pengujian model	100
3	26-07-2021	data train	pengujian dan jurnal	100
4	28-07-2021	model fix	pengujian	100
5	04-08-2021	metode word embedding fasttext dan word2vec	train model with word embedding	100
6	06-08-2021	metode word embedding	train model with word embedding fasttext vs word2vec vs glove	100
7	07-08-2021	progress word embedding	perbandingan 3 model word embedding jurnal	100
8	10-08-2021	perbandingan word embedding	deploy flask dan visualisasi tableau	100
9	11-08-2021	pengujian cnn	deploy flask dan visualisasi tableau	100
10	12-08-2021	pengujian fix	visualisasi tableau dan finishing	100
			Rata-Rata:	100.00

Bandung, 18 Agustus 2021

Pembimbing,



Rolly Maulana Awangga S.T., M.T.



POLITEKNIK POS INDONESIA  
PROGRAM STUDI D4 TEKNIK INFORMATIKA  
JL. SARIASIH NO. 54 BANDUNG 40151  
Telp. 022-2009562, 2009570  
Fax. 022-2009568

**FORMULIR KEGIATAN**  
**TUGAS AKHIR**  
**TA. 2020/2021**

Nama : Muhammad Fahmi  
Npm : 1174021  
Judul : Analisis Sentimen Vaksin Covid-19 Berdasarkan  
Lokasi Di Indonesia Menggunakan Metode Lexicon  
Based Dan CNN  
Pembimbing : Roni Andarsyah ST., M.KOM



Pertemuan	Tanggal	Sudah Dikerjakan	Pekerjaan Selanjutnya	Nilai
1	14-07-2021	bab 1	revisi bab 1 kerjakan bab 2	90
2	21-07-2021	progress bab 1 2	bab 3	100
3	28-07-2021	bab 3	modelling dan mulai pembuatan jurnal	100
4	03-08-2021	akurasi dan confusion matrix	akurasi modelling cnn dan lanjut pembuatan jurnal	100
5	05-08-2021	akurasi cnn	modelling word embedding dan lanjut pembuatan jurnal	100
6	07-08-2021	progress word embedding	perbandingan 3 model word embedding	100
7	11-08-2021	perbandingan word embedding	modelling dan pengujian cnn	100
8	12-08-2021	proses modelling 6000 data	testing 28000 data	100
9	14-08-2021	proses pengujian cnn	finishing keseluruhan	100
10	17-08-2021	finishing keseluruhan	submit jurnal	100
			Rata-Rata:	99.00

Bandung, 18 Agustus 2021

Pembimbing,



Roni Andarsyah ST., M.KOM

## Lampiran 2. Bukti Pengecekan Plagiarisme Online

TA Fahmi

### ORIGINALITY REPORT

19%

SIMILARITY INDEX

18%

INTERNET SOURCES

5%

PUBLICATIONS

8%

STUDENT PAPERS

### PRIMARY SOURCES

1

[ejurnal.teknokrat.ac.id](http://ejurnal.teknokrat.ac.id)

Internet Source

4%

2

[repositori.unsil.ac.id](http://repositori.unsil.ac.id)

Internet Source

1%

3

[eprints.ums.ac.id](http://eprints.ums.ac.id)

Internet Source

1%

4

[repository.its.ac.id](http://repository.its.ac.id)

Internet Source

1%

5

Submitted to Universitas Brawijaya

Student Paper

1%

6

Submitted to IAIN Kudus

Student Paper

1%

7

[jurnal.uns.ac.id](http://jurnal.uns.ac.id)

Internet Source

1%



### Lampiran 3. Surat Pertanyaan

#### **SURAT PERNYATAAN TIDAK MELAKUKAN PLAGIARISME**

Yang bertanda tangan di bawah ini :

Nama : Muhammad Fahmi

NPM : 1174021

Program Studi : D4 Teknik Informatika

Judul : ANALISIS SENTIMENVAKSIN COVID-19 BERDASARKAN  
LOKASI DI INDONESIA MENGGUNAKAN METODE  
*LEXICON BASED* DAN *CNN*.

Menyatakan bahwa :

1. Program tugas akhir saya ini adalah asli dan belum pernah diajukan untuk memenuhi kelulusan matakuliah tugas akhir pada Program Studi D4 Teknik Informatika baik di Politeknik Pos Indonesia maupun di Perguruan Tinggi lainnya.
2. Program tugas akhir ini adalah murni gagasan, rumusan dan penelitian saya sendiri tanpa bantuan pihak lain, kecuali arahan pembimbing.
3. Dalam Program tugas akhir ini tidak terdapat karya atau pendapat yang telah ditulis atau dipublikasikan orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan dicantumkan dalam daftar pustaka.
4. Pertanyaan ini saya buat dengan sesungguhnya dan apabila di kemudian hari terdapat penyimpangan-penyimpangan dan ketidakbeneran dan pernyataan ini, maka saya bersedia menerima sanksi akademik berupa pencabutan gelar yang telah diperoleh karena karya ini, serta sanksi lainnya sesuai dengan norma yang berlaku diperguruan tinggi lain.

Bandung, 18 Agustus 2021



Muhammad Fahmi  
NPM. 1.17.4.021

#### Lampiran 4. Bukti Submit Jurnal

**Songklanakar Journal of Science and Technology**  
**Sentiment Analysis Of Twitter Users On Covid-19 Vaccine By Location In Indonesian**  
**Using Lexicon Based and CNN**  
--Manuscript Draft--

<b>Manuscript Number:</b>	
<b>Full Title:</b>	Sentiment Analysis Of Twitter Users On Covid-19 Vaccine By Location In Indonesian Using Lexicon Based and CNN
<b>Article Type:</b>	Original Article
<b>Section/Category:</b>	Engineering
<b>Keywords:</b>	Sentiment; Vaccine; Lexicon Based; CNN; FastText
<b>Corresponding Author:</b>	Rolly Maulana Awangga Politeknik Pos Indonesia Bandung, Jawa Barat INDONESIA
<b>Corresponding Author Secondary Information:</b>	
<b>Corresponding Author's Institution:</b>	Politeknik Pos Indonesia
<b>Corresponding Author's Secondary Institution:</b>	
<b>First Author:</b>	Muhammad Fahmi
<b>First Author Secondary Information:</b>	
<b>Order of Authors:</b>	Muhammad Fahmi Rolly Maulana Awangga Roni Andarsyah
<b>Order of Authors Secondary Information:</b>	
<b>Manuscript Region of Origin:</b>	INDONESIA
<b>Abstract:</b>	On its official website, the outbreak of the Covid-19 virus disease was designated as a global pandemic by WHO on March 11, 2020. So one way to prevent the spread of this virus is to develop a vaccine. However, many opinions are given by the public on Twitter social media. Therefore, the author will analyze Twitter users' positive, negative, and neutral sentiments about the Covid-19 vaccine on Twitter social media based on the distribution of provinces in Indonesia. This research uses Lexicon Based and CNN (Convolutional Neural Network) methods from Deep Learning Python. Thus, a conclusion can be drawn in the form of dashboard visualization. This study will compare several word embeddings, including FastText Word2vec and GloVe. The highest accuracy obtained on the CNN+FastText model is 97%.
<b>Suggested Reviewers:</b>	Yusril Helmi Setyawan Politeknik Pos Indonesia yusrilhelmi@poltekpos.ac.id
<b>Opposed Reviewers:</b>	
<b>Additional Information:</b>	
<b>Question</b>	<b>Response</b>
Please enter the <b>Word Count</b> of your manuscript	3776

