ANALISIS PENINGKATAN AKURASI METODE DISTILBERT DALAM MENGKLASIFIKASI *TWEET* MENGENAI COVID-19

OLEH :

FAISAL FAJRI

09012681923004

PROGRAM STUDI MAGISTER ILMU KOMPUTER
FAKULTAS ILMU KOMPUTER
UNIVERSITAS SRIWIJAYA

TAHUN 2023

ANALISIS PENINGKATAN AKURASI METODE DISTILBERT DALAM MENGKLASIFIKASI *TWEET* MENGENAI COVID-19

TESIS

Diajukan Untuk Melengkapi Salah Satu Syarat
Memperoleh Gelar Magister

OLEH :

FAISAL FAJRI

09012681923004

PROGRAM STUDI MAGISTER ILMU KOMPUTER
FAKULTAS ILMU KOMPUTER
UNIVERSITAS SRIWIJAYA

TAHUN 2023

LEMBAR PENGESAHAN

ANALISIS PENINGKATAN AKURASI METODE DISTILBERT DALAM MENGKLASIFIKASI *TWEET* MENGENAI COVID-19

TESIS

Diajukan untuk Melengkapi Salah Satu Syarat
Memperoleh Gelar Magister

OLEH:

FAISAL FAJRI
09012681923004

Palembang, 20 Maret 2023

Pembimbing I

21/3/2023

Dr. Ir. Bambang Tutuko, M.T.
NIP. 196001121989031002

Pembimbing II

20/3/23

Dr. Ir. Sukemi, M.T.
NIP. 196612032006041001

Mengetahui,
Koordinator Program Studi Magister Ilmu Komputer

Hadipurnawan Satria, Ph.D.
NIP. 198004182020121001

HALAMAN PERSETUJUAN

Pada hari Jumat, 21 Desember 2022 telah dilaksanakan ujian sidang tesis oleh Magister Ilmu Komputer Fakultas Ilmu Komputer Universitas Sriwijaya.

Nama : Faisal Fajri
NIM : 09012681923004
Judul : Analisis Peningkatan Akurasi Metode DistilBERT dalam Mengklasifikasi Tweet Mengenai Covid-19

1. Pembimbing 1

   Dr. Ir. Bambang Tutuko, M.T. .....
   NIP. 196001121989031002

2. Pembimbing II

   Dr. Ir. Sukemi, M.T. .....
   NIP. 196612032006041001

3. Penguji I

   Dr. Iwan Pahendra AS, S.T., M.T. .....
   NIP. 197403222002121002

4. Penguji II

   Dr. Yusuf Hartono, M.sc. .....
   NIP. 196411161990031002

Mengetahui,
Koordinator Program Studi Magister Ilmu Komputer

Hadipurnawan Satria, Ph.D.
NIP. 198004182020121001

LEMBAR PERNYATAAN

Yang bertanda tangan di bawah ini :

Nama	: Faisal Fajri
NIM	: 09012681923004
Program Studi	: Magister Ilmu Komputer
Judul Tesis	: Analisis Peningkatan Akurasi Metode DistilBERT Dalam Mengklasifikasi Tweet Mengenai Covid-19

Hasil Pengecekan Software iThenticate/Turnitin : **2 %**

Menyatakan bahwa laporan tesis saya merupakan hasil karya sendiri dan bukan hasil penjiplakan/plagiat. Apabila ditemukan unsur penjiplakan/plagiat dalam laporan tesis ini, maka saya bersedia menerima sanksi akademik dari Universitas Sriwijaya sesuai dengan ketentuan yang berlaku.

Demikian, pernyataan ini saya buat dengan sebenarnya dan tidak ada paksaan oleh siapapun.

Palembang, Maret 2023

Faisal Fajri

NIM. 09012681923004

KATA PENGANTAR

Segala puja dan puji syukur penulis panjatkan kehadirat Allah SWT karena atas berkat dan rahmat yang telah diberikan oleh-Nya, sehingga penulis dapat menyelesaikan Tesis dengan judul "**Analisis Peningkatan Akurasi Metode DistilBERT Dalam Mengklasifikasi Tweet Mengenai Covid-19**", sebagai salah satu persyaratan yang wajib dipenuhi oleh mahasiswa Universitas Sriwijaya, yang berguna untuk memperoleh gelar Magister Ilmu Komputer. Tak lupa juga penulis panjatkan Shalawat serta salam bagi junjungan Nabi Muhammad SAW yang menjadikan pedoman bagi kehidupan umat muslim di seluruh dunia.

Penulis menyampaikan banyak terimakasih yang sebesar-besarnya terhadap berbagai kalangan serta pihak yang telah memberikan dorongan dalam menyusun penulisan skripsi ini, antara lain:

1. Orang tua yang penulis sayangi, Ayahanda H. Darwin, A.md. dan Ibunda Hj. Uliana yang telah memberikan dukungan, motivasi dan doa restu kepada Ananda sehingga bisa menyelesaikan tesis ini.
2. Istri tercinta Bettaria, S.SI yang telah mensupport dan mendoakan penulis untuk menyelesaikan tesis, walaupun banyak halangan dan rintangan yang penulis dan istri lalui selama studi dan tesis ini.
3. Adik kandung tersayang dr. Melly Ratna Sari dan Ramadhan Putra, S.T. serta adik ipar Dr. Baldi Anggara, M.Pd.I yang selalu mensupport dan mendoakan penulis agar dapat menyelesaikan tesis tepat waktu.
4. Dosen pembimbing Bapak Dr. Ir. Bambang Tutuko, M.T. dan Dr. Ir. Sukemi, M.T. yang selalu memberikan motivasi dan bimbingan selama dalam proses penulisan tesis dan jurnal sehingga dapat berlangsung secara baik dan benar.
5. Dr. Yusuf Hartono, M.Sc. dan Dr. Iwan Pahendra AS, S.T., M.T. selaku dosen penguji sidang tesis yang telah memberikan masukan dan saran dalam penelitian tesis agar menjadi lebih baik lagi.
6. Seluruh dosen Fakultas Ilmu Komputer pada umumnya dan dosen Program Studi Magister Ilmu Komputer pada khususnya yang telah memberikan ilmunya dan memberikan masukan serta saran kepada penulis selama masa studi penulis di lingkungan Fakultas Ilmu Komputer Universitas Sriwijaya ini.

7. Ardina Ariani, M.Kom. selaku admin Program Studi Magister Ilmu Komputer yang telah banyak membantu penulis dalam memperlancar kegiatan perkuliahan sampai sidang tesis ini selesai.
8. Teman-teman almamater Program Studi Magister Ilmu Komputer Universitas Sriwijaya yang telah membantu penulis selama masa perkuliahan sampai sidang tesis ini selesai.
9. Rekan kerja di PT. Telkom Indonesia unit RWS yang selalu mensupport dan mendoakan penulis agar bisa secepatnya menyelesaikan studi S2, terutama pimpinan Ibu Febriza Matillya S.R, M.T. yang telah memberikan izin kepada penulis untuk melanjutkan studi S2 di Fakultas Ilmu Komputer Universitas Sriwijaya.
10. Semua pihak baik yang secara langsung maupun tidak langsung telah membantu penulis untuk menyelesaikan studi S2 di Fakultas Ilmu Komputer Universitas Sriwijaya.

Penulis menyadari bahwa selama penyelesaian tesis dan menimba ilmu di Program Studi Magister Ilmu Komputer Universitas Sriwijaya ini, penulis masih sangat banyak sekali kekurangan. Sehingga saran, kritik serta dukungan dari teman-teman sangat membantu penulis dalam penyelesaian karya tulis khususnya yang berkenaan dengan penelitian dalam Tesis ini. Akhir kata, penulis sangat berharap sekali penulisan karya ilmiah berupa Tesis ini bisa bermanfaat bagi semua pihak, khususnya Program Studi Magister Ilmu Komputer Universitas Sriwijaya.

Palembang, Maret 2023

Penulis

ANALYSIS OF IMPROVING THE ACCURACY OF THE DISTILBERT METHOD IN CLASSIFYING TWEETS ABOUT COVID-19

Faisal Fajri (09012681923004)
Dept of Master Computer Science, Computer Science Faculty, Sriwijaya University
Email : faisal.fajri88@gmail.com

ABSTRACT

Sentiment analysis is a fundamental task in Natural Language Processing (NLP). Social media is designed to enable people to share content quickly through electronic tools. People can openly express their minded on social media sites like Twitter, which later can be shared with others. During the recent COVID-19 outbreak, public opinion analytics provided useful information for determining the best public health response. In this study, researchers will improve BERT accuracy by using the DistilBERT method. The DistilBERT classification method is designed to reduce the size and increase the training speed of the two way encoder representation of the transformer model (BERT). The experimental results using the BERT method generate an accuracy value of 87%, while using the DistilBERT method increased the accuracy value by 10%, so that the accuracy value using the DistilBERT method becomes 97%.

Keywords: *Coronavirus* – 19, *Tweet*, BERT, DistilBERT

ANALISIS PENINGKATAN AKURASI METODE DISTILBERT DALAM MENGKLASIFIKASI *TWEET* MENGENAI COVID-19

Faisal Fajri (09012681923004)
Jurusan Magister Ilmu Komputer, Fakultas Ilmu Komputer, Universitas Sriwjaya
Email : faisal.fajri88@gmail.com

ABSTRAK

Analisis sentimen adalah tugas mendasar dalam *Natural Language Processing* (NLP). Orang dapat secara terbuka mengungkapkan pemikirannya di situs media sosial seperti *Twitter*, yang kemudian dapat dibagikan kepada orang lain. Selama wabah COVID-19 baru-baru ini, analitik opini publik memberikan informasi yang berguna untuk menentukan respons kesehatan masyarakat yang terbaik. Dalam penelitian kali ini, peneliti akan meningkatkan akurasi BERT dengan menggunakan metode DistilBERT. Metode klasifikasi DistilBERT dirancang untuk mengurangi ukuran dan meningkatkan kecepatan pelatihan representasi enkoder dua arah dari model *transformer* (BERT). Hasil percobaan dengan menggunakan metode BERT menghasilkan nilai akurasi sebesar 87%, sedangkan dengan menggunakan metode DistilBERT mengalami peningkatan nilai akurasi sebesar 10%, sehingga nilai akurasi dengan menggunakan metode DistilBERT menjadi 97%.

Kata kunci: *Coronavirus* – 19, *Tweet*, BERT, DistilBERT

DAFTAR ISI

LEMBAR PENGESAHAN iii

HALAMAN PERSETUJUAN iv

LEMBAR PERNYATAAN v

KATA PENGANTAR vi

ABSTRACT viii

ABSTRAK ix

DAFTAR TABEL xiii

DAFTAR GAMBAR xiv

BAB I 1

PENDAHULUAN 1

1.1 Latar Belakang 1

1.2 Perumusan Masalah 7

1.3 Batasan Masalah 8

1.4 Tujuan Penelitian 8

1.5 Manfaat Penelitian 8

1.6 Metodologi Penelitian 9

BAB II 11

TINJAUAN PUSTAKA 11

2.1 Tinjauan Penelitian 11

2.2 Atribut *Dataset Coronavirus Tweet* 14

2.3 *Sentiment Analysis* 15

2.4 *Preprocessing* 16

2.5 *Corpus* dan *Vocab* 20

2.6 *Confusion Matrix* 20

2.6.1 Akurasi 22
2.6.2 *Precision* 22
2.6.3 *Recall* 22
2.6.4 *F1-Score* 22
2.7 BERT (*Bidirectional Encoder Representations From Transformer*) 23
2.8 DistilBERT (*Distillation* BERT) 26
2.9 *Transformer Encoder* dan *Decoder* 29
2.9.1 *Encoder – Decoder Architecture* 31
BAB III 34
METODOLOGI PENELITIAN 34
3.1 Tahapan Penelitian 34
3.2 Persiapan Data 36
3.3 Alat dan Bahan Penelitian 36
3.4 Data *Twitter, Cleaning* dan Analisa Proses 36
3.5 *Preprocessing* 38
3.6 Metode DistilBERT 40
3.7 Proses Pengujian 41
3.8 Kesimpulan dan Saran 42
BAB IV 43
HASIL DAN PEMBAHASAN 43
4.1 Hasil *Preprocessing* 43
4.2 Hasil Pengujian DistilBERT 44
4.3 Hasil Pengujian *Confusion Matrix* 44
4.3.1 Perhitungan *Confusion Matrix* dengan Menggunakan Sistem 45
4.3.2 Perhitungan *Confusion Matrix* dengan Metode Manual 46
4.3.2.1 Hasil *Extremely Negative* 46

4.3.2.2 Hasil *Negative* 48

4.3.2.3 Hasil *Neutral* 50

4.3.2.4 Hasil *Positive* 51

4.3.2.5 Hasil *Extremely Positive* 53

4.3.3 Hasil Perhitungan *Confusion Matrix* dengan Menggunakan Perhitungan Secara Manual dan Secara Sistem dengan Metode DistilBERT 55

4.3.4 Perbandingan Hasil BERT dan DistilBERT 56

BAB V 58

KESIMPULAN DAN SARAN 58

5.1 Kesimpulan 58

5.2 Saran 59

DAFTAR PUSTAKA 60

DAFTAR TABEL

Tabel 2.1 *Dataset Twitter* Covid-19 15

Tabel 2.2 Tabel *Confusion Matrix* 21

Tabel 2.3 Model BERT 24

Tabel 3.1 Hasil *Preprocessing* 39

Tabel 4.1 Hasil *Preprocessing* 43

Tabel 4.2 Hasil Percobaan Metode DistilBERT dengan Menggunakan *learning rate* 3e-05 dan 4 *epoch* 44

Tabel 4.3 Hasil Perhitungan *Confusion Matrix* dengan Menggunakan Sistem ... 45

Tabel 4.4 *Sentiment Analysis* DistilBERT 46

Tabel 4.5 *Confusion Matrix Extremely Negative* 46

Tabel 4.6 Hasil perhitungan manual *confusion matrix Extremely Negative* 48

Tabel 4.7 *Confusion Matrix Negative* 48

Tabel 4.8 Hasil perhitungan manual *confusion matrix Negative* 49

Tabel 4.9 *Confusion Matrix Neutral* 50

Tabel 4.10 Hasil perhitungan manual *confusion matrix Neutral* 51

Tabel 4.11 *Confusion Matrix Positive* 51

Tabel 4.12 Hasil perhitungan manual *confusion matrix Positive* 53

Tabel 4.13 *Confusion Matrix Extremely Positive* 53

Tabel 4.14 Hasil perhitungan manual *confusion matrix Extremely Positive* 54

Tabel 4.15 Hasil *Confusion Matrix* 55

Tabel 4.16 Perbandingan Hasil BERT dan DistilBERT 56

DAFTAR GAMBAR

Gambar 2.1 Diagram Alir *Corpus* dan *Vocab* 20

Gambar 2.2 BERT *Framework* 25

Gambar 2.3 Model Ekstraksi Fitur 28

Gambar 2.4 Model Arsitektur *Transformer* 31

Gambar 2.5 The *Transformer* 31

Gambar 2.6 *Encoder* 32

Gambar 2.7 *Decoder* 32

Gambar 2.8 Arsitektur *Transformer Encoder-Decoder* 37

Gambar 3.1 Diagram Alir Tahapan Penelitian 34

Gambar 3.2 Diagram Alir *Sentiment Analysis* dengan *Dataset Twitter* 37

Gambar 3.3 Arsitektur dan Komponen DistilBERT 41

Gambar 4.1 Matrik 5x5 *Confusion Matrik* DistilBERT 45

BAB I PENDAHULUAN

Pada bab ini menjelaskan tentang latar belakang penelitian yang berjudul "Analisis Peningkatan Akurasi Metode DistilBERT dalam Mengklasifikasi Tweet Mengenai Covid-19". Mengapa penulis mengangkat topik ini, karena covid-19 penyakit baru yang telah menjadi pandemi. Penyakit ini harus diwaspadai karena penularan yang relatif cepat, memiliki tingkat mortalitas yang tidak dapat diabaikan. Oleh karena itu, segala pembahasan mengenai *tweet* covid-19 akan menjadi info yang sangat berguna bagi masyarakat luas.

1.1 Latar Belakang

Wabah penyakit baru yang disebabkan oleh virus korona (2019-nCoV) atau yang biasa disebut dengan COVID-19 ditetapkan secara resmi sebagai pandemi global oleh *World Health Organization* (WHO) pada tanggal 11 Maret 2020 lalu. Penyakit *Coronavirus* 2019 (COVID-19) adalah penyakit menular yang sangat menular dengan implikasi kesehatan global yang besar. Meskipun pusat penyebaran virus tersebut pada akhir tahun 2019 lalu berada di Kota Wuhan, China, kini virus tersebut telah tersebar menjangkit ke seluruh masyarakat dunia dengan jumlah kasus sebanyak lebih dari 41,5 juta kasus dan jumlah kematian sebanyak lebih dari 1,1 juta jiwa per tanggal 23 Oktober 2020. Pada 31 Januari 2021, ada 103 juta infeksi yang dikonfirmasi di seluruh dunia, merenggut lebih dari 2,2 juta jiwa. Rintangan utama dalam pengelolaan dan pengendalian COVID-19 adalah ketersediaan tes *skrining* dan pemantauan penyakit yang tepat waktu (Shakouri, et al. 2021). *Coronavirus* adalah virus RNA dengan ukuran partikel 120-160 nm. Virus ini utamanya menginfeksi hewan, termasuk di antaranya adalah kelelawar dan unta. Sebelum terjadinya wabah COVID-19, ada 6 jenis *coronavirus* yang dapat menginfeksi manusia, yaitu *alphacoronavirus* 229E, *alphacoronavirus* NL63, *betacoronavirus* OC43, *betacoronavirus* HKU1, *Severe Acute Respiratory Illness Coronavirus* (SARS-CoV), dan *Middle East Respiratory Syndrome Coronavirus* (MERS-CoV) (Susilo, A., et al. 2020). Kondisi demikian memberikan dampak langsung kepada jutaan bahkan seluruh masyarakat dunia, sebagai akibat dari

diberlakukannya protokol kesehatan yang harus ditetapkan pada seluruh aspek kegiatan, mulai dari pembatasan sosial hingga *lockdown* total sehinga menghambat seluruh kegiatan masyarakat. Efek lanjutan dari COVID-19 ini berpotensi membawa tantangan besar bagi sistem kesehatan dunia dan memiliki konsekuensi yang luas pada ekonomi *global* jika penyebaran virus tidak dikendalikan secara efektif.

Melihat pesatnya penyebaran COVID-19 dan bahaya yang akan muncul jika tidak segera ditangani, salah satu cara yang sangat mungkin untuk mencegah penyebaran virus ini adalah dengan mengembangkan vaksin (Liu C, et al. 2020). Vaksin tidak hanya melindungi mereka yang divaksinasi tetapi juga masyarakat luas dengan mengurangi penyebaran penyakit dalam populasi (Sari IP, Sriwidodo. 2020). Meskipun tidak ada vaksin untuk SARS dan MERS yang ditemukan, vaksin COVID-19 dapat ditemukan terlebih dahulu. Pengembangan vaksin yang aman dan efektif sangat penting dilakukan karena diharapkan dapat menghentikan penyebaran dan mencegah penyebaran penyakit di masa mendatang (Liu C, et al. 2020). Hal ini mempengaruhi banyak sektor kehidupan masyarakat. Tak sedikit masyarakat yang aktif bersosial media dan menuliskan pendapat, opini serta pemikirannya di *platform* media sosial seperti *Twitter*. Terjadinya pandemi ini mendorong masyarakat untuk menuliskan opini, pemikiran serta pendapatnya terhadap COVID-19 pada media sosial *Twitter*. Dibutuhkan suatu model *sentiment analysis* untuk mengklasifikasi *tweet* masyarakat di *Twitter* menjadi positif dan negatif (Fairuz, A. L., Ramadhani, R. D., & Tanjung, N. A. F. 2021). *Platform* media sosial memainkan peran yang lebih penting secara *global* daripada sebelumnya (Naseem, U., Razzak, I., & Eklund, P. W. 2021). Analisis informasi yang dibagikan di *platform* media sosial, khususnya *Twitter*, telah menjadi fokus yang signifikan bagi para peneliti dalam beberapa tahun terakhir.

Jutaan pengguna *Twitter* membagikan pendapat dan pandangan mereka tentang berbagai topic, seperti debat politik, pasar saham, produk, perusahaan, dan sebagainya terutama mengenai permasalah yang sedang hangat yaitu mengenai pandemi COVID-19. Pesan *Twitter* dibatasi hingga 140 karakter, sehingga bahasa yang digunakan di *Twitter* dinormalisasi dengan batasan ini, yaitu tidak terstruktur, dan terkadang sangat informal. Pertumbuhan dunia media digital tumbuh dengan

kecepatan yang luar biasa, yang membuat konsumsi informasi menjadi tugas yang menantang. Karena *volume* data teks *online* yang terus meningkat, tugas klasifikasi teks menjadi lebih penting dari sebelumnya. Dalam konteks ini, klasifikasi teks (secara otomatis mengklasifikasikan tekstual) adalah tugas yang penting. Sebagian besar media digital dibuat oleh pengguna, tetapi mencari informasi yang diperlukan secara manual berada di luar kemampuan manusia (Gao, Z., Feng, A., Song, X., & Wu, X. 2019). Pemrosesan media sosial yang dibantu oleh *Machine Learning* sangat membantu dalam era yang serba digital. *Natural Language Processing* (NLP) atau biasa disebut juga pemrosesan bahasa alami adalah teori yang termotivasi dari teknik komputasi untuk analisa otomatis dan representasi bahasa manusia.

NLP memungkinkan komputer untuk melakukan tugas bahasa alami, seperti penguraian dan pelabelan kelas kata, terjemahan mesin, hingga yang populer saat ini adalah analisis sentimen. Analisis sentimen adalah tugas mendasar dalam *Natural Language Processing* (NLP). *Aspect-Based Sentiment Analysis* (ABSA), adalah tugas terperinci dalam analisis sentimen, yang bertujuan untuk mengidentifikasi polaritas sentimen (misalnya, positif, negatif, netral, konflik) dari kategori aspek atau target (juga disebut istilah aspek) (Pontiki, et al. 2016). Tujuan utama klasifikasi teks adalah untuk mengekstrak informasi dari sumber daya tekstual. Tugas klasifikasi teks adalah modul dasar untuk banyak aplikasi NLP, namun hal ini memerlukan adanya metode yang efisien dan fleksibel untuk mengakses, mengatur, dan mengekstrak informasi yang berguna dari sumber data yang berbeda. Metode-metode ini dapat mencakup klasifikasi teks, pencarian informasi, peringkasan, pengelompokan teks, dan lain-lain yang secara kolektif dinamakan penambangan teks.

Perkembangan teknologi saat ini tumbuh sangat pesat, hal ini membuat penyebaran informasi semakin mudah dan cepat melalui media *online* (*facebook, twitter*), blog, atau situs resmi suatu lembaga (Huddar, M. G., et al, 2021). Dengan kemudahan dan kecepatan media *online*, mampu mengubah cara konsumsi masyarakat terhadap suatu berita. Berdasarkan banyaknya pengguna media sosial tersebut maka jumlah data yang tersimpan di media sosial juga semakin banyak. Sehingga para peneliti banyak melakukan penelitian dengan memanfaatkan data

media sosial ini. *Opinion mining* (OM) atau *Sentiment Analysis* (SA) merupakan salah satu metode analisa data media sosial yang mengolah sebuah informasi yang terkandung dalam teks. OM/SA ini merupakan cabang ilmu dari *Text Mining* (Nurrohmat, M. A., & SN, A. 2019). Dalam penelitian-peneilitian terdahulu, telah banyak sekali contoh penelitian yang membahas tentang *sentiment analysis*. Terdapat bermacam-macam data yang digunakan dalam implementasi dalam penyelesaian task terhadap analisis sentimen tersebut. Macam-macam data yang digunakan adalah *movie review*, teks hadits, dan bahkan komentar di sosial media (Putri, C. A. 2020).

Salah satu teknik pemrosesan bahasa alami yang sedang populer yaitu dengan menggunakan metode BERT (*Bidirectional Encoder Representations from Transformer*). BERT adalah makalah terbaru yang diterbitkan oleh para peneliti di Google *Artificial Intelengence* (AI) Language. BERT telah menyebabkan kegemparan di komunitas pembelajaran mesin dengan memberikan hasil yang mutakhir dalam berbagai tugas NLP, termasuk *Question Answering* (SQuAD v1.1), *Natural Language Inference* (MNLI), dan lainnya. BERT merupakan algoritma *Deep Learning* (DL) yang dirancang untuk mengolah NLP. Model BERT telah menunjukkan kinerja mutakhir pada banyak tugas, dan arsitektur transformernya yang dalam adalah tipikal dari banyak model terbaru. BERT dirancang untuk melatih representasi dua arah yang mendalam dari teks yang tidak berlabel dengan mengkondisikan bersama pada konteks kiri dan kanan di semua lapisan (J. Devlin. et al, 2019). Sejauh ini, baru terdapat sedikit penelitian terkait dengan analisis sentimen dengan menggunakan algoritma BERT yang dilakukan *fine-tuning* dengan beberapa *layer*.

BERT pertama kali diteliti dengan judul BERT: *Pre-training of Deep Bidirectional Transformers for Language Understanding* (J. Devlin. et al, 2019) dimana pada penelitian tersebut BERT mampu memperoleh hasil yang mutakhir pada sebelas tugas pemrosesan bahasa alami yang dilakukan. Dimana GLUE *Score* menghasilkan nilai sebesar 80,5% (mengalami peningkatan sebesar 7,7%), akurasi MultiNLI menjadi 86,7% (mengalami peningkatan sebesar 4,6%), SQuAD v1.1 *question answering* Test F1 menjadi 93,2 (peningkatan sebesar 1,5 poin) dan SQuAD v2.0 *Test* F1 menjadi 83,1 (peningkatan sebesar 5,1 poin).

(Geetha dan D. Karthika Renuka, 2021) melakukan pengembangan penelitian BERT dengan menggunakan TF-IDF yang merupakan struktur blok standar untuk beberapa algoritma pembelajaran mesin, lalu menggunakan klasifikasi *Naive Bayes* (NB) dan *Support Vector Machine* (SVM) sehingga menghasilkan nilai akurasi sebesar 87,69%, nilai *recall* sebesar 85,98%, *precision* 83,22% dan *F1 Measure* sebesar 88,72% pada *learning rate* 5e-05. *Naive Bayes* mempunyai keuntungan mampu mengklasifikasi sejumlah kecil data pelatihan, *Naive Bayes* juga mampu memahami ulasan positif dan negative. SVM merupakan klasifikasi yang bersifat *non-probabilitas*, SVM hanya menggunakan dua jenis label yaitu 0 dan 1. Lalu penelitian yang dilakukan oleh (Phuc Do dan Truong H.V. Phan, 2021) menggunakan teknik BERT-*based triple classification* mampu menghasilkan nilai akurasi secara signifikan sebesar 92,34%. Teknik ini dilatih menggunakan metode tiga rangkap, setiap 3 rangkap jalur (h, p, t) mengubah *text* dengan menggunakan deskripsi dari entitas h, entitas *tail*, dan *predicate*.

Dari penelitian sebelumnya mengenai BERT, peneliti akan melakukan peningkatan akurasi pada BERT dengan menggunakan metode yang disebut DistilBERT. Alasan mengapa peneliti menggunakan metode DistilBERT karena metode ini lebih kecil, lebih cepat, lebih murah dan lebih ringan. Maka dari itu peneliti mencoba membuktikan apakah dengan menggunakan metode DistilBERT bisa mendapatkan hasil yang di harapkan. Ada beberapa penelitian yang menguatkan peneliti untuk membuktikan apakah metode DistilBERT mampu menghasilkan nilai akurasi yang lebih tinggi dengan kemampuan proses *fine tuning* sekitar 40% lebih cepat dan mempertahankan kemampuan akurasi yang lebih tinggi bila dibandingkan dengan penggunaan metode BERT. Pada penelitian yang dilakukan oleh (Rafael Silva Barbon dan Ademar Takeo Akabane, 2022) dengan judul Towards Transfer Learning Techniques - BERT, DistilBERT, BERTimbau, and DistilBERTimbau for Automatic Text Classification from Different Languages: A Case Study, dimana metode DistilBERT dan DistilBERTimbau mampu menghasilkan model sekitar 40% lebih kecil dan membutuhkan waktu sekitar 45% (dimana percobaan yang dihasilkan berkisar antara 21,5% hingga 66,9%) lebih sedikit untuk proses *fine tuning.* Dengan kata lain, model kompresi membutuhkan sumber daya komputasi yang rendah dan dengan kinerja yang lebih besar. Sementara itu, model yang disuling

mempertahankan sekitar 96% kemampuan pemahaman bahasa untuk set data yang yang seimbang.

Penelitian yang dilakukan oleh (Berfu Buyukoz, Ali Hurriyetoglu, Arzucan Ozgur, 2020) dengan judul *Analyzing ELMo and DistilBERT on Socio-political News Classification* dimana dalam penelitian ini, ELMo dan DistilBERT dibandingkan dalam hal kinerja *fine tuning* pada dua tugas klasifikasi teks biner. Fokus utama dari penelitian ini adalah untuk melihat seberapa besar manfaat dari kedua *modifier* ini secara praktis tanpa melakukan modifikasi pada *output* pra pelatihan. Secara keseluruhan, DistilBERT ternyata dapat menggeneralisasi lebih baik daripada ELMo pada pengaturan lintas konteks. DistilBERT mengungguli ELMo dalam persentase penurunan skor F. Selain itu, DistilBERT ternyata 30% lebih kecil dalam ukuran penyematan dan 83% lebih cepat dalam waktu pelatihan bila dibandingkan dengan ELMo.

Lalu peneilitan yang dilakukan oleh (Mario Jojoa, et al. 2022) *Natural language processing analysis applied to COVID-19 open-text opinions using a distilBERT model for sentiment categorization.* Dimana penelitian yang dilakukan adalah mengurangi struktur kompleks dengan jutaan parameter pada metode BERT tetapi dengan versi yang lebih sederhana dan berkapasitas tinggi. Apakah versi distilasi dari BERT yang asli, tetapi dengan jumlah parameter yang lebih sedikit, memungkinkan untuk menyempurnakan model dalam waktu singkat dan dengan sumber daya perangkat keras menengah. Dimana peneliti telah mengusulkan model transformator DistilBERT untuk melaksanakan tugas ini. Peneliti telah menggunakan tiga pendekatan untuk melakukan perbandingan, dan memperoleh metrik rata-rata berikut untuk model terbaik dengan nilai Akurasi sebesar 0,823, Presisi : 0,826, Recall : 0.793 dan F1 Score : 0.803.

Dalam penelitian ini akan dibahas mengenai sentimen analisis masyarakat terhadap pandemi COVID-19 pada media sosial *Twitter*. Berdasarkan peristiwa yang saat ini sedang ramai di masyarakat, banyak pengguna media sosial yang memberikan opini, pendapat, serta pemikirannya terhadap COVID-19 pada *platform* media sosial *Twitter*. Hal ini menarik untuk diteliti guna mengetahui opini masyarakat tentang pandemi yang sedang terjadi sekarang ini. Untuk menunjang penelitian tentang hal tersebut, dibutuhkan algoritma untuk mengklasifikasikan

komentar masyarakat di media sosial *Twitter*, baik itu komentar positif maupun komentar negatif, metode yang akan digunakan adalah DistilBERT atau *Distillation* BERT. Peneliti akan melakukan komparasi terhadap performa algoritma *Deep Learning* yaitu antara BERT dan DistilBERT dan memfokuskan peningkatan akurasi dengan metode DistilBERT.

Dari penelitian yang telah dilakukan oleh para peneliti mengenai DistilBERT, maka peneliti akan mencoba meneliti apakah penggunaan metode DistilBERT akan mampu menghasilkan nilai akurasi yang tinggi bila dibandingkan dengan metode BERT. Dataset yang akan digunakan pada penelitian kali ini ialah *Coronavirus tweets* NLP - *Text Classification* dataset ini memberikan informasi *tweet* mengenai isu seputar *coronavirus* yang ditulis oleh para pengguna *twitter*. Karena penyebaran virus yang cepat, organisasi kesehatan dunia menyatakan keadaan darurat. Dalam tesis kali ini, peneliti akan coba memaksimalkan penggunaan metode pada DistilBERT dalam klasifikasi teks mengenai Covid-19 dari *twitter* untuk mengungkap berbagai isu terkait Covid-19 dari opini publik. Dari hasil yang didapatkan, apakah metode yang digunakan layak untuk digunakan pada klasifikasi teks, yang berdasarkan dari nilai akurasi, *recall*, *precission*, dan *F1-score* yang didapatkan.

1.2 Perumusan Masalah

Pada latar belakang yang telah dijelaskan sebelumnya terdapat beberapa isu yang akan dibahas dalam penelitian sesuai dengan penjelasan dari latar belakang di atas. Maka dari itu perlu perumusan beberapa masalah dalam penelitian kali ini, yaitu:

1. Apakah penggunaan metode DistilBERT dalam mengklasifikasi *tweet* mengenai *Covid*-19 mampu menghasilkan nilai akurasi yang tinggi bila dibandingkan dengan penggunaan metode BERT?
2. Perbandingan nilai akurasi, *recall*, *precission*, dan *F1-score* untuk klasifikasi teks dengan menggunakan metode BERT dan metode DistilBERT
3. Bagaimana hasil analisis dari peningkatan akurasi dengan menggunakan metode DistilBERT?

1.3 Batasan Masalah

Ada beberapa batasan masalah yang ditentukan dalam tesis ini, yaitu:

1. Data yang akan digunakan merupakan dataset dari Kaggle, yaitu *Coronavirus tweets* NLP – *Text Classification.*
2. Metode yang akan digunakan yaitu BERT dan DistilBERT.
3. Acuan yang menjadi perbandingan hanya akurasi, *recall*, *precission*, dan *F1-score*.

1.4 Tujuan Penelitian

Adapun tujuan yang ingin dicapai pada penelitian ini adalah:

1. Sebagai salah satu pilihan penggunaan teknik klasifikasi teks
2. Penggunaan metode DistilBERT sebagai salah satu teknik klasifikasi teks untuk meningkatkan nilai akurasi.
3. Dengan penggunaan metode DistilBERT diharapkan bisa mendapatkan peningkatan pada nilai akurasi, *recall*, *precission*, dan *F1-score* dalam klasifikasi teks.

1.5 Manfaat Penelitian

Hasil yang akan didapatkan dari penelitian kali ini adalah:

1. Memberikan nilai tambah bagi penelitian pada bidang klasifikasi teks.
2. Memberikan alternatif penggunaan metode DistilBERT sebagai salah satu solusi peningkatan nilai akurasi pada klasifikasi teks.
3. Penelitian yang dilakukan dapat meningkatkan pemahaman peneliti mengenai teknik klasifikasi teks dengan menggunakan metode DistilBERT.

1.6 Metodologi Penulisan

Agar memperoleh gambaran jelas mengenai penelitian ini, maka dibuatlah suatu sistematika penulisan yang berisi gambaran dalam tiap bab penelitian ini, yaitu:

1. **BAB I PENDAHULUAN**

   Bab ini menjelaskan tentang latar belakang, perumusan masalah, batasan masalah, tujuan dan manfaat dari topik yang dipilih berupa peningkatan akurasi pada metode DistilBERT dalam klasifikasi teks.

2. **BAB II TINJAUAN PUSTAKA**

   Bab ini menjelaskan mengenai *literature review* peningkatan klasifikasi metode DistilBERT dengan menggunakan dataset *tweet* Covid-19 yang mengacu pada penelitian publikasi mengenai metode DistilBERT. Algoritma yang akan digunakan dalam penelitian, dan penambahan beberapa metode yang bisa mendukung peningkatan nilai akurasi.

3. **BAB III METODOLOGI PENELITIAN**

   Pada bab ini penulis melakukan pembahasan secara bertahap dan terperinci langkah-langkah yang akan digunakan pada penelitian mengenai metode peningkatan akurasi BERT yaitu DistilBERT. Parameter yang akan digunakan untuk menentukan hasil pengujian yaitu akurasi, *recall*, *precision*, dan *F1-score*.

4. **BAB IV HASIL DAN ANALISA**

   Pada bab hasil dan analisa, berisi mengenai hasil pengujian yang telah dilakukan. Data-data hasil pengujian akan diolah dengan perhitungan *Sentiment Analysis* sehingga bisa didapatkan nilai akurasi, *recall*, *precision*, dan *F1-score.*

5. BAB V KESIMPULAN

Bab ini merupakan kesimpulan dari penelitian yang telah dilakukan, dan bisa mendapatkan jawaban apakah metode yang digunakan mengalami peningkatan nilai akurasi pada klasifikasi teks.

DAFTAR PUSTAKA

Gimpel, K., Schneider, N., Connor, B. O., Das, D., Mills, D., Eisenstein, J., … Smith, N. A. (2011). *Part-of-Speech Tagging for Twitter : Annotation , Features , and Experiments*. 42–47.

Chitraa, V. (2010). *A Survey on Preprocessing Methods for Web Usage Data*. *7*(3), 78–83.

Gao, Z., Feng, A., Song, X., & Wu, X. (2019). Target-dependent sentiment classification with BERT. *IEEE Access*, *7*, 154290–154299. https://doi.org/10.1109/ACCESS.2019.2946594

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Al-Smadi, M., … Eryigit, G. (2016). SemEval-2016 task 5: Aspect based sentiment analysis. *SemEval 2016 - 10th International Workshop on Semantic Evaluation, Proceedings*, 19–30. https://doi.org/10.18653/v1/s16-1002

Huddar, M. G., Sannakki, S. S., & Rajpurohit, V. S. (2021). Attention-based multimodal contextual fusion for sentiment and emotion classification using bidirectional LSTM. *Multimedia Tools and Applications*, *80*(9), 13059–13076. https://doi.org/10.1007/s11042-020-10285-x

Nurrohmat, M. A., & SN, A. (2019). Sentiment Analysis of Novel Review Using Long Short-Term Memory Method. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *13*(3), 209. https://doi.org/10.22146/ijccs.41236

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*(Mlm), 4171–4186.

Geetha, M. P., & Renuka, D. K. (2021). International Journal of Intelligent Networks Improving the performance of aspect based sentiment analysis using fi ne-tune Bert Base Uncased model. *International Journal of Intelligent Networks*, *2*(March), 64–69. https://doi.org/10.1016/j.ijin.2021.06.005

Do, P., & Phan, T. H. V. (2021). Developing a BERT based triple classification model using knowledge graph embedding for question answering system. *Applied Intelligence*. https://doi.org/10.1007/s10489-021-02460-w

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. 2–6. Retrieved from http://arxiv.org/abs/1910.01108

Naseem, U., Razzak, I., & Eklund, P. W. (2021). A survey of pre-processing techniques to improve short-text quality: a case study on hate speech detection on twitter. *Multimedia Tools and Applications*, *80*(28–29), 35239–35266. https://doi.org/10.1007/s11042-020-10082-6

Hermanto, D. T., Setyanto, A., & Luthfi, E. T. (2021). Algoritma LSTM-CNN untuk Binary Klasifikasi dengan Word2vec pada Media Online. *Creative Information Technology Journal*, *8*(1), 64. https://doi.org/10.24076/citec.2021v8i1.264

Dwi, R., Santosa, W., Bijaksana, M. A., & Romadhony, A. (2021). Implementasi Algoritma Long Short-Term Memory (LSTM) untuk Mendeteksi Penggunaan Kalimat Abusive Pada Teks Bahasa Indonesia. *Jurnal Tugas Akhir Fakultas Informatika*, *8*(1), 691–702.

Acheampong, F. A., Nunoo-Mensah, H., & Chen, W. (2021). Transformer models for text-based emotion detection: a review of BERT-based approaches. In *Artificial Intelligence Review* (Vol. 54). https://doi.org/10.1007/s10462-021-09958-2

Adel, H., Dahou, A., Mabrouk, A., Elaziz, M. A., Kayed, M., El-Henawy, I. M., … Ali, A. A. (2022). Improving Crisis Events Detection Using DistilBERT with Hunger Games Search Algorithm. *Mathematics*, *10*(3). https://doi.org/10.3390/math10030447

Faturrohman, F., & Rosmala, D. (n.d.). *Analisis Sentimen Sosial Media dengan Metode Bidirectional Gated Recurrent Unit*. *X*(X), 1–10.

Basiri, M. E., Nemati, S., Abdar, M., Asadi, S., & Acharrya, U. R. (2021). A novel fusion-based deep learning model for sentiment analysis of COVID-19 tweets. *Knowledge-Based Systems*, *228*, 107242. https://doi.org/10.1016/j.knosys.2021.107242

Ayoub, J., Yang, X. J., & Zhou, F. (2021). Combat COVID-19 infodemic using explainable natural language processing models. *Information Processing and Management*, *58*(4), 102569. https://doi.org/10.1016/j.ipm.2021.102569

Mater, A., & Universit, S. (2019). *Deep Question Answering : A New Teacher For DistilBERT*.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). Bag of tricks for efficient text classification. *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference*, *2*, 427–431. https://doi.org/10.18653/v1/e17-2068

Liu C, Zhou Q, Li Y, Garner L V, Watkins SP, Carter LJ, et al. Research and Development on Therapeutic Agents and Vaccines for COVID-19 and Related Human Coronavirus Diseases. 2020

Sari IP, Sriwidodo. Perkembangan Teknologi Terkini dalam Mempercepat Produksi Vaksin Covid-19. 2020;5(5):204–17

Fairuz, A. L., Ramadhani, R. D., & Tanjung, N. A. F. (2021). Analisis Sentimen Masyarakat Terhadap COVID-19 Pada Media Sosial Twitter. *Journal of Dinda : Data Science, Information Technology, and Data Analytics*, *1*(1), 42–51. https://doi.org/10.20895/dinda.v1i1.180

Putri, C. A. (2020). Analisis Sentimen Review Film Berbahasa Inggris Dengan Pendekatan Bidirectional Encoder Representations from Transformers. *JATISI (Jurnal Teknik Informatika Dan Sistem Informasi)*, *6*(2), 181–193. https://doi.org/10.35957/jatisi.v6i2.206

Büyüköz, B., Hürriyetoğlu, A., & Özgür, A. (2020). Analyzing ELMo and DistilBERT on Socio-political News Classification. *Proceedings of the Workshop on Automated Extraction of Socio-Political Events from News 2020*, pp. 9–18. Retrieved from https://aclanthology.org/2020.aespen-1.4%0Ahttps://www.aclweb.org/anthology/2020.aespen-1.4

Jojoa, M., Eftekhar, P., Nowrouzi-Kia, B., & Garcia-Zapirain, B. (2022). Natural language processing analysis applied to COVID-19 open-text opinions using a distilBERT model for sentiment categorization. *AI and Society*, (0123456789). https://doi.org/10.1007/s00146-022-01594-w

Komang, I., Ganda Wiguna, A., Sugiartawan, P., Gede, I., Sudipa, I., Putu, I., & Pratama, Y. (2022). Sentiment Analysis Using Backpropagation Method to Recognize the Public Opinion. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems*, *16*(4), 423–434. Retrieved from https://doi.org/10.22146/ijccs.78664

Barbon, R. S. (2021). *Classification from Different Languages : A Case Study.*