KINERJA METODE *CONVOLUTIONAL NEURAL NETWORK* (CNN) DAN *LONG-SHORT TERM MEMORY* (LSTM) PADA KLASIFIKASI DATA JUDUL BERITA COVID-19

(Skripsi)

Oleh

LUTHFIA NUR AZIZAH

1817031065

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG

2022

ABSTRACT

PERFORMANCE OF CONVOLUTIONAL NEURAL NETWORK (CNN) AND LONG-SHORT TERM MEMORY (LSTM) METHODS ON DATA CLASSIFICATION OF COVID-19 NEWS TITLE

By

LUTHFIA NUR AZIZAH

The performance of deep learning methods on the classification of text data with different imbalance ratios is an important discussion because the existing data is inherently imbalanced. This study is looking for a reliable deep learning method to classify data on Indonesian news headlines about COVID-19 with several data imbalance ratios. Several data imbalance ratios were made by taking samples from news events using simple random sampling of 30%, 20%, 10%, and 1%. The performance of the CNN and LSTM methods was tested using 10-fold cross validation and compared based on accuracy, precision, recall, and f1-score. The CNN model architecture built in this study generally consists of an input layer, word embedding layer, two convolutional layers, one pooling layer, flatten, two hidden layers, an output layer. A batch normalization layer and a dropout layer after each layer. The LSTM model architecture built in this study generally consists of an input layer, a word embedding layer, two LSTM layers, two hidden layers, an output layer and a dropout layer after each layer. The performance of CNN and LSTM with the Bag of Words (BoW) model as word embedding in this study is quite competitive because CNN outperforms LSTM on all evaluation measures at 37%, 20%, and 10% data imbalance levels, while LSTM outperforms CNN on all evaluation measures at 30% data imbalance levels. Although CNN and LSTM have competitive performance results, LSTM consumes significantly longer computational time than CNN.

Keywords: Classification, Imbalance Data, Deep Learning, CNN, LSTM, K-Fold Cross Validation, Bag of Words (BoW), COVID-19 news headlines.

ABSTRAK

KINERJA METODE *CONVOLUTIONAL NEURAL NETWORK* (CNN) DAN *LONG-SHORT TERM MEMORY* (LSTM) PADA KLASIFIKASI DATA JUDUL BERITA COVID-19

Oleh

LUTHFIA NUR AZIZAH

Kinerja metode *deep learning* pada klasifikasi data teks dengan rasio ketidakseimbangan yang berbeda merupakan diskusi yang penting karena data yang ada pada dasarnya tidak seimbang. Penelitian ini mencari metode *deep learning* yang dapat diandalkan untuk mengklasifikasi data judul berita berbahasa Indonesia tentang COVID-19 dengan beberapa rasio ketidakseimbangan data. Beberapa rasio ketidakseimbangan data dibuat dengan mengambil sampel dari berita *event* menggunakan *simple random sampling* sebanyak 30%, 20%, 10%, dan 1%. Kinerja metode CNN dan LSTM diuji menggunakan 10-*fold cross validation* dan dibandingkan berdasarkan akurasi, presisi, *recall*, dan *f1-score*. Arsitektur model CNN yang dibangun pada penelitian ini secara umum terdiri dari *input layer*, *word embedding layer*, dua *convolutional layer*, satu *pooling layer*, *flatten*, dua *hidden layer*, *output layer* serta *batch normalization layer* dan *dropout layer* berada setelahnya pada setiap *layer* tersebut. Arsitektur model LSTM yang dibangun pada penelitian ini secara umum terdiri dari *input layer*, *word embedding layer*, dua *LSTM layer*, dua *hidden layer*, *output layer* serta *dropout layer* berada setelahnya pada setiap *layer* tersebut. Kinerja CNN dan LSTM dengan model *Bag of Words* (BoW) sebagai *word embedding* pada penelitian ini cukup bersaing karena CNN mengungguli LSTM pada semua ukuran evaluasi pada tingkat ketidakseimbangan data 37%, 20%, dan 10%, sedangkan LSTM mengungguli CNN pada semua ukuran evaluasi pada tingkat ketidakseimbangan data 30%. Meskipun CNN dan LSTM memiliki hasil kinerja yang saling bersaing, namun LSTM menghabiskan waktu komputasi yang jauh lebih lama daripada CNN.

Kata kunci: Klasifikasi, Ketidakseimbangan Data, *Deep Learning*, CNN, LSTM, *K-Fold Cross Validation*, *Bag of Words* (BoW), Judul Berita COVID-19.

KINERJA METODE *CONVOLUTIONAL NEURAL NETWORK* (CNN) DAN *LONG-SHORT TERM MEMORY* (LSTM) PADA KLASIFIKASI DATA JUDUL BERITA COVID-19

Oleh

LUTHFIA NUR AZIZAH

Skripsi

Sebagai Salah Satu Syarat untuk Memperoleh Gelar
SARJANA MATEMATIKA

Pada

Jurusan Matematika
Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Lampung

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2022

Judul Skripsi	:	**KINERJA METODE *CONVOLUTIONAL NEURAL NETWORK* (CNN) DAN *LONG-SHORT TERM MEMORY* (LSTM) PADA KLASIFIKASI DATA JUDUL BERITA COVID-19**
Nama Mahasiswa	:	Luthfia Nur Azizah
Nomor Pokok Mahasiswa	:	**1817031065**
Jurusan	:	**Matematika**
Fakultas	:	**Matematika dan Ilmu Pengetahuan Alam**

MENYETUJUI

1. Komisi Pembimbing

Dian Kurniasari, S.Si., M.Sc.	**Dr. Purnomo Husnul Khotimah, M.T.**
NIP. 196903051996032001	NIP. 198003232005022002

2. Ketua Jurusan Matematika

Dr. Aang Nuryaman, S.Si., M.Si.
NIP. 197403162005011001

MENGESAHKAN

1. Tim Penguji

   Ketua : **Dian Kurniasari, S.Si., M.Sc.** .....

   Sekretaris : **Dr. Purnomo Husnul Khotimah, M.T.** .....

   Penguji
   Bukan Pembimbing : **Ir. Warsono, M.S., Ph.D.** .....

2. Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

Dr. Eng Suripto Dwi Yuwono, S.Si., M.T.
NIP. 19740705 200003 1 001

Tanggal Lulus Ujian Skripsi : **05 Agustus 2022**

PERNYATAAN SKRIPSI MAHASISWA

Yang bertanda tangan di bawah ini:

Nama	: Luthfia Nur Azizah
Nomor Pokok Mahasiswa	: 1817031065
Jurusan	: Matematika
Judul Skripsi	: **Kinerja Metode *Convolutional Neural Network* (CNN) dan *Long-Short Term Memory* (LSTM) pada Klasifikasi Data Judul Berita COVID-19**

Dengan ini menyatakan bahwa skripsi ini adalah hasil pekerjaan saya sendiri, bukan hasil orang lain. Apabila kemudian hari terbukti bahwa skripsi ini merupakan hasil salinan atau dibuat oleh orang lain, maka saya bersedia menerima sanksi sesuai dengan ketentuan akademik yang berlaku.

Bandar Lampung, 06 Agustus 2022

Penulis

Luthfia Nur Azizah

NPM. 1817031065

RIWAYAT HIDUP

Luthfia Nur Azizah dilahirkan di Kota Bekasi pada 29 Oktober 2000. Penulis merupakan anak pertama dari tiga bersaudara dari pasangan Ayah Alfan Choiri NST dan Ibu Erma.

Penulis menempuh pendidikan pertamanya di Taman Kanak-Kanak (TK) Fitri dan melanjutkan pendidikan Sekolah Dasar (SD) di SDN Cimuning III pada tahun 2006 – 2012. Selanjutnya, penulis melanjutkan jenjang pendidikannya di MTs Attaqwa Pusat Putri Bekasi pada tahun 2012 – 2015 dan MA Attaqwa Pusat Putri Bekasi pada tahun 2015 – 2018. Pada tahun 2018, penulis diterima sebagai mahasiswa Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lampung melalui jalur Seleksi Bersama Masuk Perguruan Tinggi Negeri (SBMPTN).

Selama menjadi mahasiswa, penulis aktif di Komunitas Mahasiswa Penghafal Al-qur'an (MPQ) Unila sebagai Anggota Bidang Akademik pada tahun 2019 – 2020 hingga menjadi Koor Bidang Akademik pada tahun 2020 – 2021. Kemudian penulis menjadi *musyrifah* di Rumah Peradaban Qur'ani (RPQ) Lampung sampai penulis menyelesaikan penulisan skripsi ini.

Pada Februari 2021, penulis melaksanakan Kuliah Kerja Nyata (KKN) di Desa Tegal Yoso, Kecamatan Purbolinggo, Kabupaten Lampung Timur. Kemudian penulis melakukan Kerja Praktik (KP) di SAMSAT Rajabasa, Bandar Lampung pada Juli 2021. Selain itu, penulis mengikuti kegiatan penelitian program Merdeka Belajar Kampus Merdeka (MBKM) di Kelompok Penelitian *Information Retrieval* (IR), Pusat Riset Informatika, Badan Riset dan Inovasi Nasional (BRIN), Bandung,

sejak bulan Oktober tahun 2021 – Februari 2022. Selama penelitian, penulis menghasilkan sebuah karya tulis ilmiah dengan judul "*The Investigation into Deep Learning Classifiers Towards Imbalanced Text Data*" yang dipresentasikan pada *International Converence* NISS 2022 dan sedang dalam proses penerbitan.

KATA INSPIRASI

“Dan bersabarlah kamu. Sesungguhnya janji Allah adalah benar.”

(Q.S Ar-Rum : 60)

“Dan hanya kepada Tuhanmulah hendaknya kamu berharap.”

(Q.S Al-Insyirah : 8)

“Sungguh atas kehendak Allah semua ini terwujud, tiada kekuatan kecuali dengan pertolongan Allah.”

(Q.S Al-Kahfi : 39)

“Seseorang yang bersabar tidak akan pernah kehilangan kesuksesan meskipun membutuhkan waktu yang lama untuk mencapainya.”

(Ali Bin Abi Thalib)

“*Lillah*.”

(Erma)

“Segala sesuatu yang dimulai dengan basmalah, semoga keberkahan menyertainya.”

(Penulis)

PERSEMBAHAN

Alhamdulillah, puji dan syukur kepada Allah SWT atas nikmat serta hidayahnya sehingga skripsi ini dapat terselesaikan dengan baik dan tepat pada waktunya.

Oleh karena itu, dengan rasa syukur dan bahagia saya persembahkan rasa terimakasih saya kepada

Ibu dan Ayah

Tidak ada kata yang dapat fia sampaikan untuk kalian kecuali ucapan terimakasih atas semua yang telah kalian berikan untuk fia. Cinta, kasih sayang, motivasi, waktu, pengorbanan yang belum bisa fia balas, serta doa dan sujud yang selalu menantikan keberhasilan fia dengan sabar dan penuh pengertian. Terimakasih karena selalu mendoakan dan mendukung setiap langkah yang fia pilih. Karena atas doa dan ridho kalian, Allah memudahkan setiap perjalanan hidup fia. Terimalah bukti kecil ini sebagai kado keseriusan fia untuk membalas semua pengorbanan, keikhlasan, dan jerih payah yang selama ini kalian lakukan.

Dosen Pembimbing dan Pembahas

Terimakasih kepada dosen pembimbing dan pembahas yang sudah sangat membantu, memberikan motivasi, memberikan arahan serta ilmu yang berharga.

Almamater Tercinta Universitas Lampung

SANWACANA

Assalamu'alaikum warahmatullahi wabarakatuh

Puji syukur saya haturkan ke hadirat Allah SWT yang telah melimpahkan nikmat, rahmat, hidayah, serta pertolongan-Nya kepada penulis sehingga skripsi ini dapat diselesaikan.

Skripsi dengan judul **"Kinerja Metode *Convolutional Neural Network* (CNN) dan *Long-Short Term Memory* (LSTM) pada Klasifikasi Data Judul Berita COVID-19"** dibuat sebagai bentuk pertanggungjawaban penulis selama menempuh pendidikan S1 dan merupakan salah satu syarat untuk memperoleh gelar Sarjana Matematika (S.Mat.) di Universitas Lampung.

Penulis menyadari bahwa selama proses penulisan skripsi ini masih jauh dari kata sempurna. Proses penyusunan skripsi ini tentu tidak luput dari pengarahan, kritik, saran, dukungan, serta bimbingan dari berbagai pihak sehingga dapat terselesaikan pada waktu yang tepat. Dalam kesempatan ini, penulis menyampaikan rasa hormat dan ucapan terima kasih kepada:

1. Ibu Dian Kurniasari S.Si., M.Sc. selaku dosen pembimbing I yang senantiasa membimbing dengan sabar, memberi masukan serta saran serta mendukung penulis dalam menyelesaikan skripsi ini.
2. Ibu Dr. Purnomo Husnul Khotimah, M.T. selaku dosen pembimbing II memberikan bimbingan, pengarahan, serta saran sehingga penulis dapat menyelesaikan skripsi ini.
3. Bapak Ir. Warsono, M.S., Ph.D. selaku dosen penguji yang telah memberikan kritik dan saran yang membangun sehingga skripsi ini dapat terselesaikan.

4. Bapak Prof. Dr. Lazakaria, S.Si., M.Sc., selaku dosen pembimbing akademik yang senantiasa memberikan saran dan bimbingan selama penulis mengemban pendidikan di bangku perkuliahan.
5. Bapak Dr. Aang Nuryaman, S.Si., M.Si., selaku Ketua Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.
6. Bapak Dr. Eng. Suripto Dwi Yuwono, S.Si., M.T. selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.
7. Seluruh dosen, staff, karyawan Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.
8. Teruntuk kedua orang tuaku tercinta, Ayah Alfan Choiri NST dan Ibu Erma terimakasih atas doa, dukungan, pengorbanan, cinta kasih, perhatian, demi kesuksesan penulis semoga dikemudian hari dapat membahagiakan dan menjadi kebanggan kalian.
9. Om, tante, kakak dan adik sepupuku, serta keluarga besar yang senantiasa menyemangati penulis untuk menggapai cita-cita.
10. Bu Diana, Bu Iin, Bu Wiwin, Pak Devi, Pak Andri, dan Pak Andria dari Kelompok Penelitian *Information Retrieval* di Pusat Riset Informatika yang telah banyak membantu, membimbing dan mengayomi selama penelitian.
11. Teruntuk sahabat ku sejak MTs sampai saat ini dan seterusnya. Terimakasih kepada Ulfi, Fitri, Adel, dan Sem Aul atas doa, dorongan, saran, dan motivasi, serta dukungan dalam menyelesaikan skripsi ini.
12. Sahabat kuliahku: Linda, Risa, Aul, Intan, Shavira, Febi, Hilda, Mona, Kintan, Dora, dan Ria, terimakasih atas pengalaman serta dukungan terhadap penulis dari sejak awal perkuliahan hingga selesai. Semoga kita terus jadi kawan baik sampe seterusnya.
13. Sahabat kost terbaik, Arum dan Restu, terima kasih banyak atas kebersamaan, kepedulian, doa, saran, dan dukungan kalian. Sukses terus !
14. Semua teman sejurusan matematika 2018 dan teman kelas B yang telah membantu serta memberikan semangat kepada penulis yang mana tidak bisa disebutkan satu persatu.
15. Teman-teman seperbimbinganku Maydia, Virda, Putsal, Sulis, Dalfa, Alip, Farrel, Ferzy, dan Zaenal, terima kasih atas doa, motivasi, dukungan, semangat,

yang selalu di berikan kepada penulis. Semoga kalian menjadi orang yang sukses dan bahagia dimanapun kalian berada. *See you on top, guys!*

16. Keluarga langit seperjuangan skripsi, Mba Valen, Mba Naja, Mba Devio, Mba Nurfi, Bella, Aini, Anggia, Mba Dian, Mba Helen, Dewi, dan Dika, terima kasih atas semua doa, dukungan, saran, dan motivasi. Semoga kalian sehat, sukses, dan berkah selalu, *Uhibbukunna fillah*.
17. Seluruh keluarga langitku, terkhusus Mba Nurul, Mba Marda, Adibah, Halida, Muthi, Mba Mely, Mba Syifa, Mba Dilah, Mba Eka, Mba Dijah, Nabilah, Yasmin, Nunik, Arfa, Rani, Mba Puput, Riri, Vinna, dan Mba Erlia, terima kasih atas semua doa, dukungan, saran, dan motivasi. Semoga kalian sehat, sukses, dan berkah selalu, *Uhibbukunna fillah*.
18. Sahabatku sejak SD sampai saat ini dan seterusnya, Riris, Amel, Icha, Ghisa, dan Nisa, terima kasih untuk tetap saling berkomunikasi walaupun jarak memisahkan kita.
19. *Special man*, Ka Misbahul Badri yang selalu memberikan doa, perhatian, dukungan, saran, menemani, dan mendengarkan segala cerita dan keluh kesah penulis. Semoga negosiasi kita diridhoi-Nya.
20. Semua pihak yang tidak dapat disebutkan satu persatu yang telah membantu penulis dalam menyelesaikan skripsi ini.

Penulis menyadari bahwa masih banyak kekurangan dalam penulisan skripsi ini. Oleh karena itu, penulis mengharapkan masukan serta saran untuk dijadikan pelajaran kedepannya.

Bandar Lampung, 06 Agustus 2022

Penulis,

Luthfia Nur Azizah

DAFTAR ISI

Halaman

DAFTAR TABEL

Tabel Halaman

DAFTAR GAMBAR

Gambar Halaman

I. PENDAHULUAN

1.1 Latar Belakang dan Masalah

Banyak Penelitian yang dilakukan untuk mencari solusi atas berbagai permasalahan dalam kehidupan sehari-hari, khususnya dalam bidang matematika dan statistika. Pemodelan matematika diperlukan untuk menerapkan matematika untuk mencari solusi dari masalah tersebut. Salah satu penelitian yang menggunakan pemodelan matematika adalah klasifikasi untuk mengetahui dinamika penyebaran penyakit menular.

Coronavirus disease 2019 (COVID-19) adalah penyakit menular yang dimulai pada akhir Desember 2019 dengan munculnya kasus pneumonia yang tidak diketahui etiologinya di Wuhan, China. COVID-19 menyebar dengan cepat di Thailand, Jepang, Korea Selatan, dan hampir di seluruh dunia hingga kasus pertama COVID-19 di Indonesia dipublikasikan pada 2 Maret 2020 (Kemenkes RI, 2020). Fenomena ini telah menyebarkan berita tentang COVID-19 ke hampir seluruh media massa, baik cetak, elektronik maupun *online* (Fadli dan Novita 2021). Kecepatan, kemudahan, dan pemberitaan yang selalu *up to date* atas segala hal yang terjadi di masyarakat menyebabkan portal berita *online* menjadi salah satu media massa yang berperan penting dalam menyebarluaskan informasi dan sering dijadikan rujukan utama oleh masyarakat (Wahyudi, dkk., 2021). Portal berita *online* juga memiliki potensi besar sebagai sumber data informal dan terpercaya untuk memantau dinamika status COVID-19 secara hampir *real-time* (Nugraheni, dkk., 2020).

Ketidakpastian munculnya varian COVID-19 menyebabkan para pemangku kepentingan seperti pemerintah, institusi kesehatan, dan masyarakat umum perlu

mewaspadai dinamika status COVID-19 agar dapat menyiapkan tindakan dan kebijakan yang diperlukan untuk mengatasi pandemi COVID-19. Dinamika status COVID-19 dapat diketahui melalui berita yang dimuat di portal berita *online*. Berita tentang COVID-19 yang dimuat tidak serta merta berarti ada kejadian COVID-19. Tidak semua berita yang diterbitkan dalam istilah umum (seperti "virus corona", "covid19", dan "pandemi") berisi topik yang relevan, beberapa diantaranya hanya berisi tentang informasi COVID-19 sehingga kurang relevan untuk memantau dinamika status COVID-19. Oleh karena itu, diperlukan metode klasifikasi yang memungkinkan para pemangku kepentingan untuk memantau dinamika status COVID-19.

Salah satu metode yang sering digunakan adalah klasifikasi teks berita namun, menemukan topik utama suatu berita dengan menganalisis seluruh isi berita akan memberatkan. Cara lain untuk menemukan topik utama suatu berita adalah dengan menganalisis judul berita. Judul berita termasuk ke dalam teks pendek, sehingga mudah untuk menemukan topik utama suatu berita berdasarkan judul berita. Fakta ini dikuatkan oleh Khotimah, dkk., (2020) yang berhasil mendeteksi wabah demam berdarah dengan mengklasifikasikan teks judul berita berbahasa Indonesia dari portal berita *online*.

Klasifikasi teks adalah metode pengelompokkan teks dalam *text mining*. Klasifikasi dalam *text mining* adalah proses pembentukan kelompok (kelas) dokumen berdasarkan kelompok (kelas) yang sudah diketahui sebelumnya (terpandu atau *supervised*) (Asiyah, 2016). Masalah yang sering muncul dalam klasifikasi adalah kelas yang tidak seimbang (*imbalanced class*) pada dataset, dimana kelas positif berjumlah lebih sedikit daripada kelas negatif. Kondisi tersebut menyebabkan pengklasifikasi cenderung salah mengklasifikasikan data yang seharusnya masuk ke dalam kelas minoritas dianggap sebagai kelas mayoritas (Hairani, dkk., 2019).

Ketidakseimbangan kelas telah menjadi subjek penelitian selama lebih dari satu dekade. Menurut L'heureux, dkk. (2017), Japkowicz dan Stephen dalam eksperimennya telah menunjukkan bahwa tingkat keparahan masalah

ketidakseimbangan kelas tergantung pada kompleksitas tugas, rasio ketidakseimbangan kelas, dan jumlah data yang digunakan pada proses pelatihan. Semakin tinggi tingkat ketidakseimbangan kelas, semakin buruk masalah bias terhadap kelas mayoritas (Liu, dkk., 2019).

Pengklasifikasi yang andal pada kasus ketidakseimbangan kelas merupakan topik penelitian yang penting (Johnson dan Khoshgoftaar, 2019), karena kelas dengan ketidakseimbangan ekstrim secara alami terlibat dalam banyak aplikasi dunia nyata seperti deteksi penipuan dan klasifikasi teks. Metode pembelajaran mendalam lebih unggul daripada metode pembelajaran mesin tradisional (seperti Naive Bayes, *Decision Tree*, *Support Vector Machine* (SVM), dan lain-lain) untuk menangani ketidakseimbangan kelas (Johnson dan Khoshgoftaar, 2019). Shaikh, dkk., pada tahun 2021 telah melakukan survei mengenai pembelajaran mendalam pada data teks (Twitter) yang sangat tidak seimbang, menyimpulkan bahwa model LSTM dengan *word embedding layer* memperoleh akurasi sebesar 81,9%. Penelitian sebelumnya yang dilakukan oleh Khotimah, dkk., (2020) menerapkan metode pembelajaran mendalam pada data judul berita berbahasa Indonesia tentang demam berdarah dengan rasio ketidakseimbangan kelas yaitu 1:2 memperoleh kesimpulan bahwa CNN dengan *word embedding layer* mampu mencapai rata-rata akurasi, presisi, *recall*, dan *f1-score* tertinggi daripada MLP dan LSTM. Diskusi mengenai kinerja metode pembelajaran mendalam pada rasio ketidakseimbangan yang berbeda dari data teks masih jarang dilakukan. Hal tersebut merupakan diskusi yang penting karena data yang ada pada dasarnya tidak seimbang. Secara umum, kelas yang penting adalah kelas minoritas, sehingga perlu diketahui metode pembelajaran mendalam mana yang paling dapat diandalkan ketika memiliki data dengan rasio ketidakseimbangan yang berbeda.

Berdasarkan pemaparan di atas, penulis tertarik untuk menemukan pengklasifikasi yang dapat diandalkan untuk data teks yang tidak seimbang. Penulis mencoba menerapkan metode pembelajaran mendalam, yaitu CNN dan LSTM menggunakan *word embedding layer* untuk membandingkan kinerja berdasarkan akurasi, presisi, *recall*, dan *f1-score* terhadap data judul berita berbahasa Indonesia tentang COVID-

19 dengan beberapa rasio ketidakseimbangan kelas (37%, 30%, 20%, 10%, dan 1%). Oleh karena itu, penulis mengangkat judul untuk penelitian ini yaitu "Kinerja Metode *Convolutional Neural Network* (CNN) dan *Long-Short Term Memory* (LSTM) pada Klasifikasi Data Judul Berita COVID-19."

1.2 Tujuan Penelitian

Adapun tujuan dari penelitian ini diantaranya:

1. Melakukan klasifikasi data judul berita berbahasa Indonesia dari portal berita *online* tentang COVID-19 menggunakan metode CNN dan LSTM dengan *word embedding layer*.
2. Membandingkan kinerja metode CNN dan LSTM dalam mengklasifikasi data teks yang tidak seimbang.

1.3 Manfaat Penelitian

Adapun manfaat penelitian ini yaitu:

1. Sebagai rujukan pengembangan ilmu matematika dalam mengklasifikasi data yang tidak seimbang serta dapat menjadi bahan pertimbangan dan informasi tambahan bagi peneliti yang akan melakukan penelitian tentang metode pembelajaran mendalam yang paling dapat diandalkan untuk data teks yang tidak seimbang.
2. Sebagai bahan pertimbangan bagi pemerintah, institusi kesehatan, masyarakat maupun pihak-pihak lain yang memiliki kepentingan terhadap pemantauan dinamika status COVID-19.

II. TINJAUAN PUSTAKA

2.1 Klasifikasi Teks

Klasifikasi teks, juga dikenal sebagai kategorisasi teks, adalah proses menetapkan label ke unit tekstual seperti kalimat, kueri, paragraf, dan dokumen berdasarkan dokumen yang telah dilabeli sebelumnya. Beberapa contoh penerapan klasifkasi teks yaitu, menjawab pertanyaan, deteksi spam, analisis sentimen, kategorisasi berita, dan sebagainya. Data teks dapat berasal dari berbagai sumber, termasuk data web, *email*, obrolan, media sosial, tiket, klaim asuransi, ulasan pengguna, dan sebagainya. Teks adalah sumber informasi yang sangat kaya, tetapi mengekstraksi wawasan dari teks bisa jadi menantang dan memakan waktu karena sifatnya yang tidak terstruktur (Minaee, dkk., 2021).

2.2 *Imbalance* Data

Ketidakseimbangan data merupakan masalah dalam klasifikasi yang telah dipelajari secara menyeluruh selama dua dekade terakhir (Johnson dan Khoshgoftaar, 2019). Ketidakseimbangan data adalah terjadinya ketidakseimbangan dalam dataset dimana salah satu kelasnya berjumlah lebih banyak (kelas mayoritas) daripada kelas lain dengan rasio bisa mencapai 1:100, 1:1000, atau bahkan 1:10000 (Ramadhanti, dkk., 2020). Kelas minoritas selalu mengandung informasi penting dalam kasus data yang tidak seimbang, sehingga model yang dibangun dari kumpulan data yang tidak seimbang sering kali bias terhadap kelas mayoritas (Santos, dkk., 2018).

2.3 *Simple Random Sampling*

Pengambilan sampel acak sederhana adalah metode pengambilan sampel yang paling sederhana. Pengambilan sampel acak sederhana adalah desain pengambilan sampel di mana *n* unit berbeda dipilih dari *N* unit dalam populasi sedemikian rupa sehingga setiap kombinasi yang mungkin dari *n* unit memiliki kemungkinan yang sama untuk terpilih sebagai sampel (Thompson, 2012).

Langkah pertama dalam pengambilan sampel acak sederhana adalah memberi setiap elemen populasi nomor 1 sampai *N*. Langkah kedua yaitu mengambil sebanyak *n* sampel dari angka-angka ini menggunakan tabel angka acak, komputer, atau kalkulator. Langkah selanjutnya adalah memastikan bahwa angka yang didapatkan berbeda. Langkah terakhir adalah menetapkan elemen populasi yang sesuai dengan angka-angka ini sebagai sampel (Levy dan Lemeshow, 2008).

2.4 *Text Mining*

Text mining merupakan proses mengekstraksi pengetahuan implisit dari data teks semi-terstruktur maupun tidak terstruktur, berbeda dengan *data mining* yang merupakan proses mengekstraksi pengetahuan implisit dari data yang sifatnya terstruktur (Rozi, dkk., 2020). Klasifikasi teks, pengelompokan, dan asosiasi adalah tugas khas dari *text mining*. *Text mining* harus dibedakan dari *information retrieval* di mana teks yang diambil adalah teks yang relevan dengan kueri. Lingkup teks pada *text mining* dibatasi pada kata-kata yang tertulis dalam bahasa alami, sedangkan kode/karakter sederhana tidak termasuk ke dalamnya (Taeho, 2019).

2.5 *Word Embedding*

Word embedding adalah salah satu pendekatan dari representasi kata dimana setiap kata dipetakan ke dalam vektor berdimensi rendah (*low-dimensional vector*) (Nurdin, dkk., 2020). Salah satu jenis *word embedding* yaitu *Bag of Words* (BoW) dengan algoritma *Count-Vectorizer.* Algoritma *Count-Vectorizer* mengubah BoW

menjadi vektor. Kalimat-kalimat dalam dokumen diekstrak ke dalam kosakata unik kemudian dihitung jumlah kemunculan setiap kata dalam setiap dokumen. Setiap dokumen diwakili oleh vektor yang ukurannya sama dengan jumlah kosakata, dan entri dalam vektor untuk dokumen tertentu menunjukkan jumlah kata dalam dokumen tersebut (Khomsah, dan Aribowo, 2020.).

2.5.1 *Word Embedding Layer*

Word embedding layer yang disediakan oleh *library* Keras merupakan *supervised learning* dengan nilai bobot yang diinisiasi secara acak dan kemudian diperbarui selama proses pelatihan model dengan menggunakan algoritma *back-propagation* (Susanty dan Sukardi, 2021).

2.6 *Deep Learning*

Deep learning merupakan salah satu cabang dari pembelajaran mesin yang pertama kali diperkenalkan pada tahun 2006 oleh Geoffrey Hinton. *Deep learning* menjadi solusi atas kekurangan dari metode pembelajaran mesin tradisional yang dapat merekayasa fitur secara otomatis atau biasa disebut dengan *feature engineering* (Rachman, 2021). Hal tersebut dapat dilakukan oleh *deep learning* karena memiliki algoritma pemodelan abstraksi tingkat tinggi pada data menggunakan sekumpulan fungsi transformasi non-linear yang ditata berlapis-lapis dan mendalam. *Deep learning* sangat baik untuk diaplikasikan pada *supervised learning*, *unsupervised learning* dan *semi-supervised learning* maupun untuk *reinforcement learning* seperti klasifikasi teks, pengenalan citra, suara, dan sebagainya (Cholissodin, dkk., 2020).

2.7 Fungsi Aktivasi

Fungsi aktivasi merupakan fungsi yang digunakan untuk mentransformasi nilai *input* menjadi nilai *output* (Brownlee, 2018). Keakuratan prediksi jaringan saraf ditentukan oleh jenis fungsi aktivasi yang digunakan. Jaringan saraf tiruan

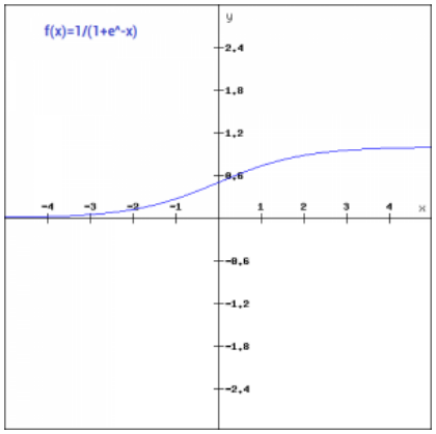
menggunakan beberapa fungsi aktivasi seperti fungsi biner, fungsi linear, fungsi sigmoid, fungsi tangen hiperbolik, fungsi ReLU, fungsi SoftMax, dan lain sebagainya (Sharma, dkk., 2017). Fungsi aktivasi yang digunakan pada penelitian ini, yaitu fungsi sigmoid, fungsi tangen hiperbolik, dan fungsi ReLU.

2.7.1 Fungsi Aktivasi Sigmoid

Fungsi sigmoid adalah fungsi non-linear yang mentransformasikan nilai dalam kisaran 0 sampai 1. Fungsi non-linear lebih disukai karena memungkinkan *node* untuk mempelajari struktur yang lebih kompleks dalam data (Brownlee, 2018). Fungsi sigmoid ditunjukkan oleh persamaan berikut:

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{2.1}$$

Berikut adalah grafik dari fungsi sigmoid:



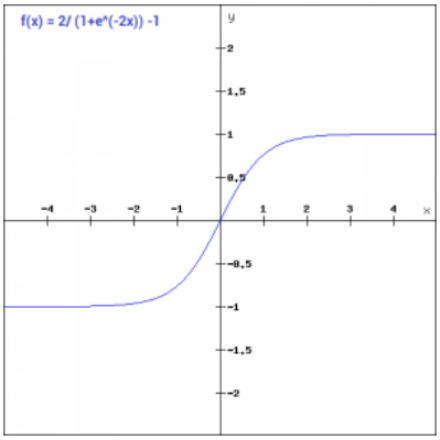
Gambar 1. Grafik Fungsi Aktivasi Sigmoid. (Sumber: Sharma, dkk., 2017).

2.7.2 Fungsi Aktivasi Tangen Hiperbolik (tanh)

Fungsi tanh juga merupakan fungsi non-linear. Jika digambarkan dalam grafik sama-sama membentuk huruf "S", namun fungsi tanh memiliki nilai dalam kisaran -1 sampai 1. Fungsi ini memiliki gradien yang lebih curam daripada fungsi sigmoid (Sharma, dkk., 2017). Fungsi tanh ditunjukkan oleh persamaan berikut:

$$\tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{2.2}$$

Berikut adalah grafik dari fungsi tanh:



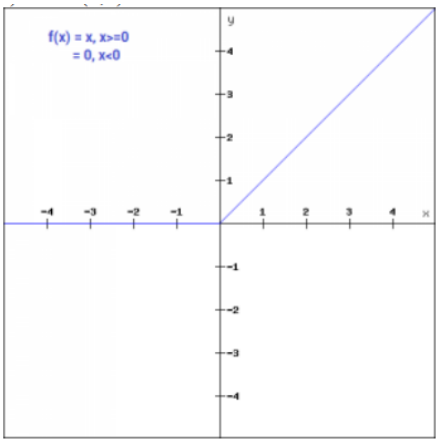
Gambar 2. Grafik Fungsi Aktivasi Tanh. (Sumber: Sharma, dkk., 2017).

2.7.3 Fungsi Aktivasi *Rectified Linear Unit* (ReLU)

Fungsi ReLU dianggap fungsi yang lebih efisien dari fungsi sigmoid dan tanh karena semua neuron tidak diaktifkan secara bersamaan, melainkan hanya neuron tertentu yang memiliki nilai *input* yang lebih besar dari 0. Artinya, jika nilai *input* lebih kecil dari 0, maka hasil *output*-nya adalah 0, ini menyebabkan neuron tersebut menjadi tidak aktif (Sharma, dkk., 2017). Fungsi ReLU ditunjukkan oleh persamaan berikut:

$$f(x) = \max(0, x) \tag{2.3}$$

Berikut adalah grafik dari fungsi ReLU:



Gambar 3. Grafik Fungsi Aktivasi ReLU. (Sumber: Sharma, dkk., 2017).

2.8 *Convolutional Neural Network* (CNN)

CNN merupakan salah satu jenis *deep learning* yang menggunakan *convolutional layer* sebagai penyusun *neural network* yang dibangun. CNN lebih baik daripada metode lain dalam mengolah data teks karena dapat menangkap konten penting dari teks dan kemampuan *convolutional layer* dalam mengurangi beban komputasi (Wang, dkk., 2021). Hal tersebut terjadi karena pada dasarnya *convolutional layer* adalah sebuah *sparse matrix* yang dimensinya lebih kecil dari dimensi data yang diolah (Widhiyasana, dkk., 2021).

Menurut Kapil, dkk. (2020), CNN adalah sejenis jaringan *deep learning* yang umum, yang terdiri dari beberapa lapisan (*layer*) yaitu:

a. *Input Layer*

   Pada *input layer*, semua urutan dikonversi ke bentuk integer dimana setiap token telah diberi indeks unik. Urutan *input* kemudian diisi nol agar memiliki panjang yang sama karena membantu meningkatkan kinerja dengan menjaga informasi tetap terjaga di perbatasan.

b. *Convolutional Layer*

   Convolutional layer pada dasarnya adalah sebuah *sparse matrix* dengan dimensi yang lebih kecil dari dimensi data yang diolah. Pada *convolutional layer* terjadi proses perkalian antara matriks *input* dengan matriks bernama *kernel* untuk menghasilkan *output* (Widhiyasana, dkk., 2021).

c. *Pooling Layer*

   Pooling layer berfungsi untuk mengurangi ukuran representasi spasial yang membantu mengurangi *overfitting*. *Max pooling* mengambil nilai maksimum lokal dari *feature map* tergantung pada ukuran *pooling*.

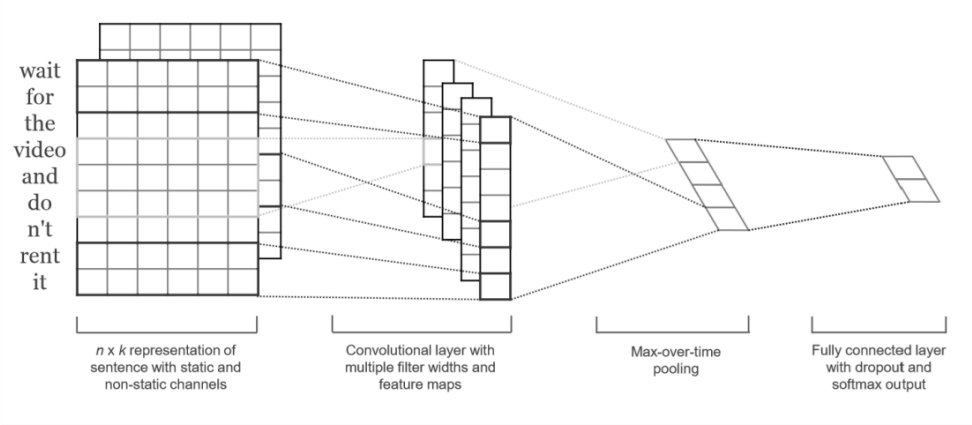
d. *Fully Connected Layer*

   Bentuk vektor dari fitur yang diperoleh dari lapisan CNN terakhir dimasukkan ke dalam *fully connected layer* yang memiliki setiap *input* yang terhubung ke setiap *output* berdasarkan bobotnya.

e. *Output layer*

Dalam klasifikasi biner, o*utput layer* pada CNN menghasilkan label berupa data numerik sebagai pemecahan masalah.

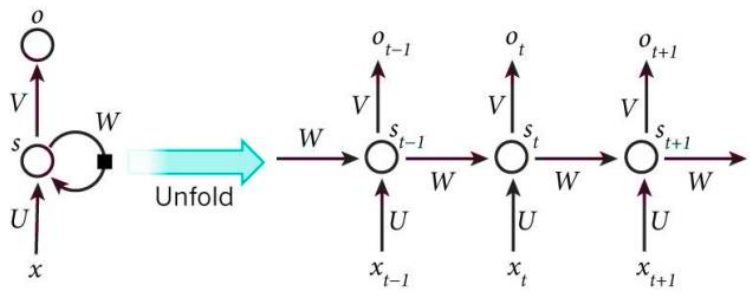
Adapun struktur CNN disajikan pada Gambar 4 berikut :



Gambar 4. Struktur CNN. (Sumber: Minaee, dkk., 2021).

2.9 *Recurrent Neural Network* (RNN)

RNN adalah bagian dari *neural network* untuk mengolah *sequential data* (Rozi, dkk., 2020). *Neural network* biasa tidak dapat memprediksi informasi *node* berikutnya berdasarkan informasi dari *node* sebelumnya, dan kata-kata dalam setiap kalimat tidak independen dalam teks. Keuntungan terbesar dari RNN adalah mereka dapat mengirimkan informasi sebelumnya ke yang berikutnya, yaitu, *output* saat ini dari suatu urutan terkait dengan *ouput* sebelumnya (Wang, dkk., 2021). Gambar 5 merupakan diagram sederhana dari struktur RNN:

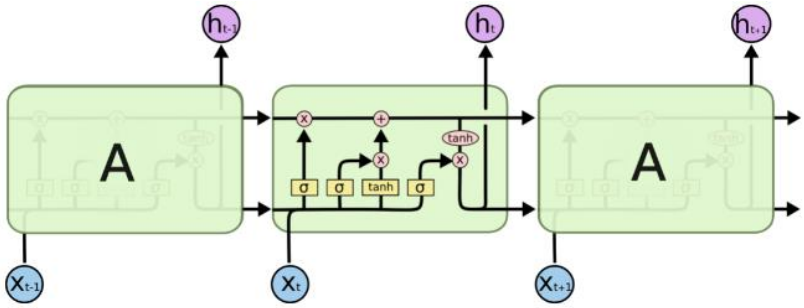


Gambar 5. Struktur RNN. (Sumber: Faturohman, dkk., 2020).

Berdasarkan gambar di atas, sisi kanan gambar adalah perluasan dari sisi kiri, dimana informasi dari sebelumnya untuk setiap *t* waktu, dikendalikan oleh bobot *W*.

2.8.1 *Long-Short Term Memory* (LSTM)

LSTM pertama kali diperkenalkan oleh Hochreiter & Schmidhuber pada tahun 1997. Jaringan LSTM menjadi solusi untuk masalah ketergantungan jangka panjang dari RNN dan merupakan jaringan yang populer untuk klasifikasi teks **(**Wang, dkk., 2021**).** LSTM memiliki model yang sama seperti RNN, tetapi struktur LSTM berbeda dengan RNN karena LSTM memiliki 4 *layer* pada setiap pengulangan seperti pada Gambar 6 berikut:



Gambar 6. *Layer* pada Setiap *Cell* LSTM. (Sumber: Olah, 2015)

LSTM mampu untuk menghapus atau menambahkan informasi ke *cell satte* yang diatur oleh struktur yang disebut *gate*. LSTM memiliki 3 *gate* untuk melindungi dan mengontrol *cell state* yaitu, *input gate*, *forget gate*, dan *output gate*. *Input gate* berfungsi untuk memperbarui nilai dari *input* pada *state memory*. *Forget gate* berfungsi untuk menentukan informasi mana yang akan dihapus dari *cell*. *Output gate* berfungsi untuk menentukan apa yang akan dihasilkan *output* sesuai dengan *input* dan memori pada *cell* (Rozi, dkk., 2020). Berikut merupakan langkah-langkah dalam LSTM:

1. *Sigmoid layer* yang bernama "*forget gate layer*" menentukan informasi apa yang akan dihapus dari *cell state*. *Forget gate layer* memproses *input* berupa h_{t-1} dan x_t, kemudian menghasilkan *output* berupa angka 0 atau 1 untuk setiap

angka dalam *cell state* C_{t-1}. Angka 1 berarti “simpan informasi ini”, sementara angka 0 mewakili “hapus informasi ini”. Adapun persamaan *forget gate layer* yaitu seperti yang ditunjukkan pada persamaan 2.4:

$$f_t = \sigma(W_f \times [h_{t-1}, x_t] + b_f) \quad (2.4)$$

dengan:

f_t : *forget gate*
W_f : nilai bobot untuk *forget gate*
x_t : nilai *input* pada orde ke *t*
σ : fungsi Sigmoid
h_{t-1} : nilai *output* sebelum orde ke *t*
b_f : nilai bias pada *forget gate*

2. Menentukan informasi baru apa yang akan disimpan di *cell state* dengan dua langkah. *Sigmoid layer* yang bernama "*input gate layer*" menentukan nilai mana yang akan diperbarui. *Tanh layer* membuat kandidat vektor nilai baru yaitu $\tilde{C}_t$, yang dapat ditambahkan ke *cell state*. *Output* dari *input gate layer* dan *tanh layer* digabungkan untuk memperbarui *cell state* (Olah, 2015). Adapun persamaan *input gate* yaitu seperti yang ditunjukkan pada persamaan 2.5:

$$i_t = \sigma(W_i \times [h_{t-1}, x_t] + b_i) \quad (2.5)$$

dengan:

i_t : *input gate*
W_i : nilai bobot untuk *input gate*
x_t : nilai *input* pada orde ke *t*
σ : fungsi Sigmoid
h_{t-1} : nilai *output* sebelum orde ke *t*
b_i : nilai bias pada *input gate*

dan persamaan kandidat vektor nilai baru ditunjukkan pada persamaan 2.6:

$$\tilde{C}_t = \tanh(W_C \times [h_{t-1}, x_t] + b_C) \quad (2.6)$$

dengan:

$\tilde{C}_t$: kandidat vektor nilai baru yang ditambahkan ke *cell state*
W_C : nilai bobot untuk *cell state*
x_t : nilai *input* pada orde ke *t*
$tanh$: fungsi tangen hiperbolik

h_{t-1} : nilai *output* sebelum orde ke *t*

b_C : nilai bias pada *cell state*

3. Memperbarui C_{t-1} (*cell state* lama) menjadi C_t (*cell state* baru), yaitu menghapus informasi yang sebelumnya sudah ditentukan dengan mengalikan *state* lama dengan f_t. Hasil perkalian tersebut ditambahkan dengan $i_t \times \tilde{C}_t$, yang merupakan kandidat vektor nilai baru untuk memperbarui setiap nilai *state*. Adapun persamaan *cell state* baru yaitu seperti yang ditunjukkan pada persamaan 2.7:

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \quad (2.7)$$

dengan:

C_t : *cell state baru*

C_{t-1} : *cell state* lama

$\tilde{C}_t$: kandidat vektor nilai baru yang ditambahkan ke *cell state*

f_t : *forget gate*

i_t : *input gate*

4. Memutuskan hasil *output*. *Output* yang dihasilkan harus sesuai dengan *cell state* yang telah diproses sebelumnya. Tahap pertama adalah *sigmoid layer* menentukan bagian mana dari *cell state* yang akan menjadi *output*. Tahap selanjutnya adalah memasukkan *output* dari *cell state* ke *tanh layer* (untuk mengganti nilai menjadi antara -1 dan 1) dan mengalikannya dengan *sigmoid gate* supaya *output* sesuai seperti yang sudah ditentukan sebelumnya. Adapun persamaan *output gate* yaitu seperti yang ditunjukkan pada persamaan 2.8:

$$o_t = \sigma(W_o \times [h_{t-1}, x_t] + b_o) \quad (2.8)$$

dengan:

o_t : *output gate*

W_o : nilai bobot untuk *output gate*

x_t : nilai *input* pada orde ke *t*

σ : fungsi Sigmoid

h_{t-1} : nilai *output* sebelum orde ke *t*

b_o : nilai bias pada *output gate*

dan persamaan nilai *output* orde ke *t* ditunjukkan pada persamaan 2.9:

$$h_t = o_t \times \tanh(C_t) \quad (2.9)$$

dengan:

h_t : nilai *output* orde ke *t*

$tanh$: fungsi tangen hiperbolik

o_t : *output gate*

C_t : *cell state*

2.10 Evaluasi Kinerja Model

Evaluasi kinerja dilakukan untuk melihat seberapa baik kinerja model klasifikasi yang digunakan. Dalam mengukur kinerja setiap model, perlu diketahui TP (*True Positive*), TN (*True Negative*), FP (*False Positive*), dan FN (*False Negative*). Nilai TP adalah jumlah kelas positif yang tepat terprediksi ke dalam kelas positif, TN adalah jumlah kelas negatif yang tepat terprediksi ke dalam kelas negatif, FP adalah jumlah kelas negatif yang terprediksi ke dalam kelas positif, dan FN adalah jumlah kelas positif yang terprediksi ke dalam kelas negatif (Kurniawan, 2017). Keempat nilai tersebut terdapat pada *confusion matrix* sebagai berikut:

Tabel 1. *Confusion Matrix*

		PREDIKSI	
		Negatif	**Positif**
AKTUAL	**Negatif**	TN	FP
	Positif	FN	TP

Ahmed dan Ahmed (2021) menggunakan akurasi, presisi, *recall*, dan *f1-score* untuk mengukur kinerja setiap model.

2.9.1 Akurasi

Akurasi secara umum adalah ukuran standar yang dapat digunakan untuk mengevaluasi kinerja model klasifikasi. Secara informal, akurasi mengacu pada persentase dari total data yang diprediksi dengan benar dalam proses klasifikasi

(Yechuri dan Ramadass, 2021). Akurasi secara matematis dinyatakan dalam persamaan 2.7:

$$\text{Akurasi} = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}} \quad (2.10)$$

2.9.2 Presisi

Presisi adalah rasio prediksi positif sejati dibandingkan dengan seluruh hasil yang diprediksi positif (Khotimah, dkk., 2020). Presisi secara matematis dinyatakan dalam persamaan 2.8:

$$\text{Presisi} = \frac{\text{TP}}{\text{TP+FP}} \quad (2.11)$$

2.9.3 *Recall*

Recall adalah rasio prediksi positif sejati dibandingkan dengan jumlah data positif yang sebenarnya (Khotimah, dkk., 2020). *Recall* secara matematis dinyatakan dalam persamaan 2.9:

$$Recall = \frac{\text{TP}}{\text{TP+FN}} \quad (2.12)$$

2.9.4 *F1-Score*

F1-score adalah rata-rata harmonik dari presisi dan *recall*. *F1-score* biasanya digunakan dalam data dengan ketidakseimbangan kelas (Khotimah, dkk., 2020). *F1-score* secara matematis dinyatakan dalam persamaan 2.10:

$$\text{F1 score} = \frac{2 \times recall \times \text{presisi}}{(recall + \text{presisi})} \quad (2.13)$$

2.11 Judul Berita *Online* COVID-19

Pandemi COVID-19 telah muncul sebagai penyakit global paling mematikan abad ini. COVID-19 yang menginfeksi hampir sepertiga populasi dunia, dan mengakibatkan 50-100 juta manusia meninggal dunia, tentunya tidak lepas dari pemberitaan di media massa khususnya portal berita *online* (Launa, 2020). Portal

berita *online* merupakan salah satu media massa yang memiliki kekuatan penting dalam menyebarluaskan informasi karena kemampuannya untuk selalu *up to date* dalam memberitakan setiap peristiwa yang terjadi di masyarakat (Wahyudi, dkk., 2021). Nugraheni, dkk. (2020) menyatakan bahwa teks berita memiliki kredibilitas yang lebih tinggi dan *noise* yang lebih rendah daripada teks media sosial, seperti *tweet*. Judul berita yang berfungsi sebagai pengantar pengetahuan pembaca tentang isi berita yang akan dideskripsikan menjadi salah satu cara untuk mengetahui isi sebuah berita (Wahyudi, dkk., 2021).

III. METODE PENELITIAN

3.1 Tempat dan Waktu Penelitian

Penelitian ini dilakukan pada semester ganjil tahun akademik 2021/2022, bertempat di Pusat Penelitian Informatika, Badan Riset dan Inovasi Nasional (BRIN), Bandung.

3.2 Spesifikasi Perangkat

Perangkat yang digunakan pada penelitian ini adalah laptop merek Lenovo dengan model IdeaPad 3 14IML05 tipe 81WA. Spesifikasi *hardware* perangkat tersebut adalah sebagai berikut:

- *Processor name* : *Intel® Pentium®* 6405U, 2 *Core(s)*
- *Processor speed* : 2400 *MHz*
- *Memory* : SSD 237 GB
- RAM : 12 GB DDR4

3.3 Data Penelitian

Pada penelitian ini digunakan data judul berita berbahasa Indonesia tentang COVID-19 yang diambil sejak bulan Januari 2020 sampai dengan Mei 2020. Data tersebut merupakan data milik Kelompok Penelitian *Information Retrieval* di Pusat Penelitian Informatika BRIN yang sudah dilabeli dengan “1” sebagai berita *event*, yaitu berita mengenai kejadian COVID-19 yang sesungguhnya dan “0” sebagai berita *non-event*, yaitu bukan berita mengenai kejadian COVID-19 yang

sesungguhnya, melainkan hanya berisi informasi COVID-19. Tabel 2 menunjukkan data yang digunakan secara rinci.

Tabel 2. Data Judul Berita COVID-19

Id	Portal Berita	Waktu Terbit	Judul Berita	Kategori Berita
778	detik	2020-01-24T11:11:00.000Z	Heboh Virus Corona yang Mirip Film 'Contagion'	0
2283	antara	2020-01-24T11:14:00.000Z	Seorang pasien "suspect" virus corona di Jakarta diisolasi	1
4449	merdeka	2020-01-24T12:00:00.000Z	Garuda Indonesia Siap Siaga Cegah Virus Corona Masuk Indonesia	0
6272	tempo	2020-01-24T14:03:00.000Z	Virus Corona Misterius: 20 Juta Warga Wuhan dan Hubei Diisolasi	1
3614	detik	2020-01-25T05:05:00.000Z	Sedih, Perayaan Imlek di China Dibatalkan akibat Wabah Corona	0
…	…	…	…	…
7571	kompas	2020-01-25T05:05:00.000Z	Virus Corona Menyebar Lewat Droplet, Kenapa Kita Perlu Cuci Tangan?	0
8112	tempo	2020-01-25T07:01:00.000Z	Bank Cina Beri Pinjaman Rp 4 triliun untuk Atasi Virus Corona	0
5657	kompas	2020-01-25T11:07:00.000Z	Xiaomi Kirim 1 Juta Masker ke Wilayah Terdampak Virus Corona	0
8648	merdeka	2020-01-25T13:30:00.000Z	Geger Virus Corona, Lion Air Alihkan 2 Rute Penerbangan Bali - Wuhan	0
5550	antara	2020-01-26T04:53:00.000Z	Data baru virus corona: 56 orang meninggal, 2.000 tertular	1

Keterangan:

Id : Identitas setiap judul berita

Portal Berita : Nama dari portal berita *online* yang menerbitkan berita tersebut, yaitu detik.com, kompas.com, antara.com, merdeka.com, republika.co.id, tempo.co, dan tirto.id

Waktu Terbit : Keterangan waktu kapan berita tersebut diterbitkan

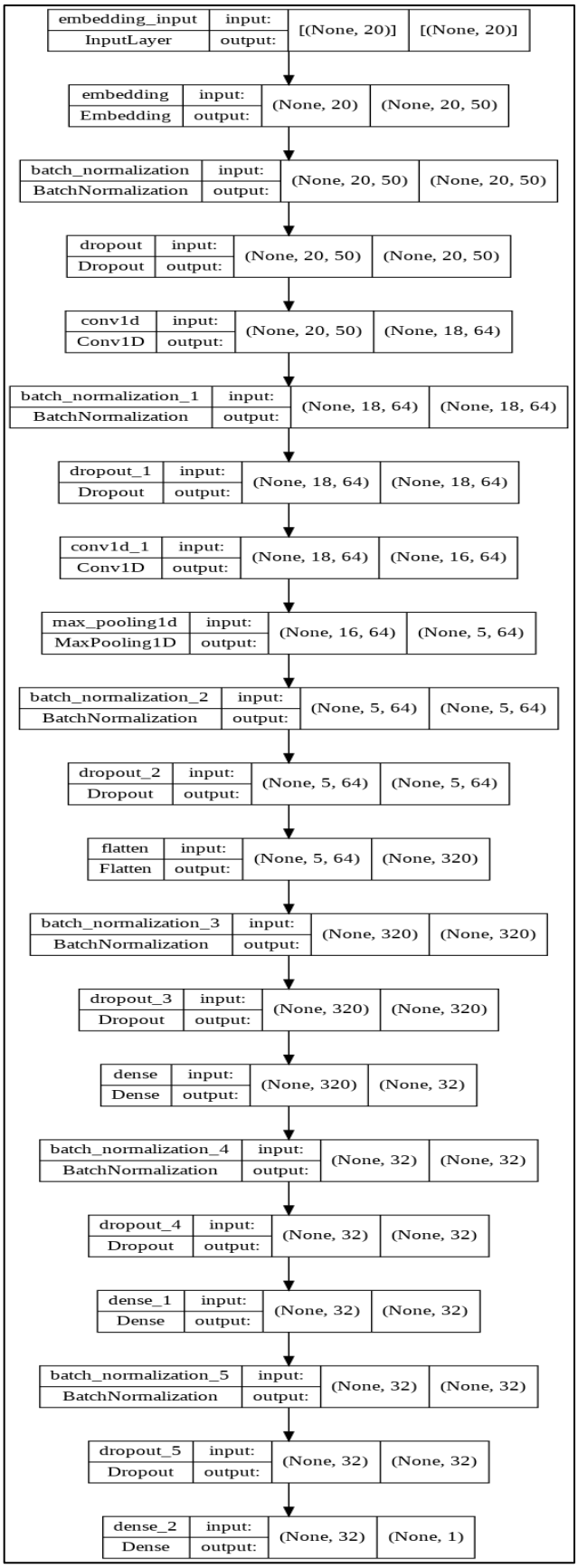
Judul Berita : Teks judul berita berbahasa Indonesia mengenai COVID-19 yang sudah diterbitkan

Kategori Berita : Label dari setiap judul berita.

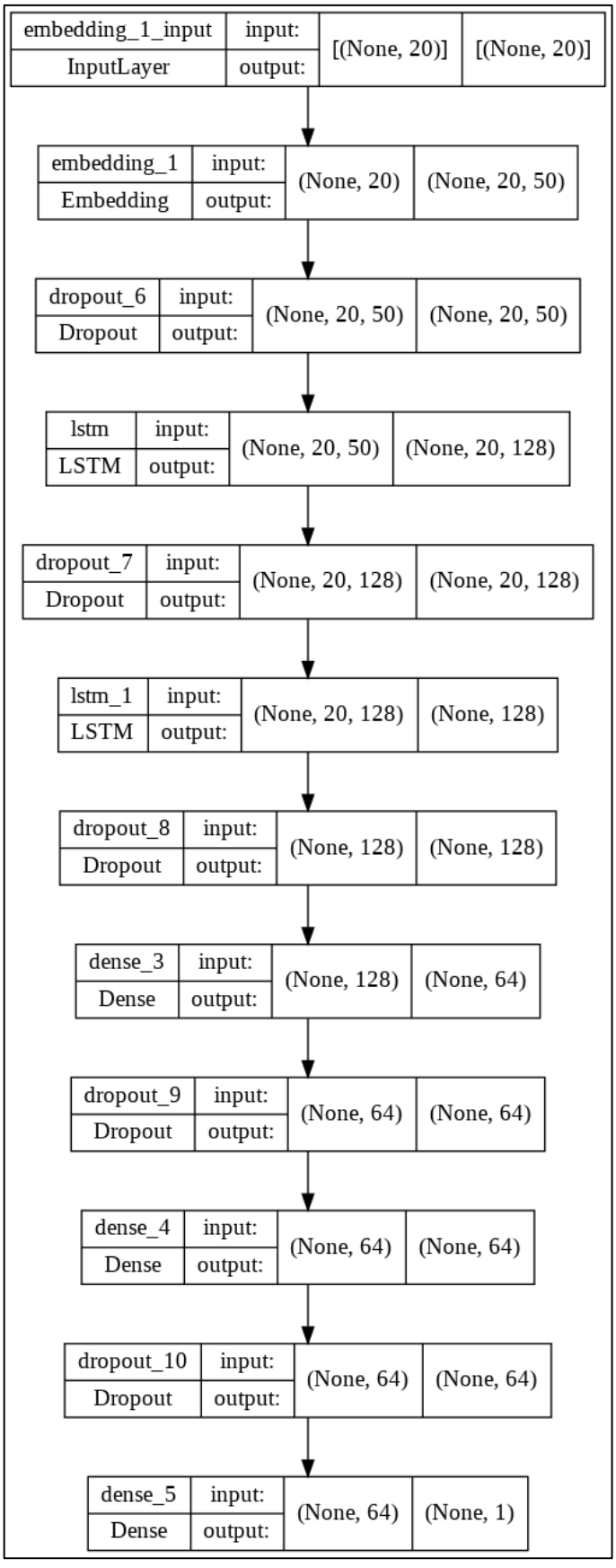
3.4 Metode Penelitian

Penelitian ini mencari model yang paling dapat diandalkan untuk mengklasifikasikan judul berita berbahasa Indonesia tentang COVID-19 pada beberapa rasio ketidakseimbangan kelas menggunakan *software* Python. Kinerja dari setiap pengklasifikasi akan dibandingkan berdasarkan akurasi, presisi, *recall*, dan *f1-score*. Adapun langkah-langkah penelitian ini yaitu:

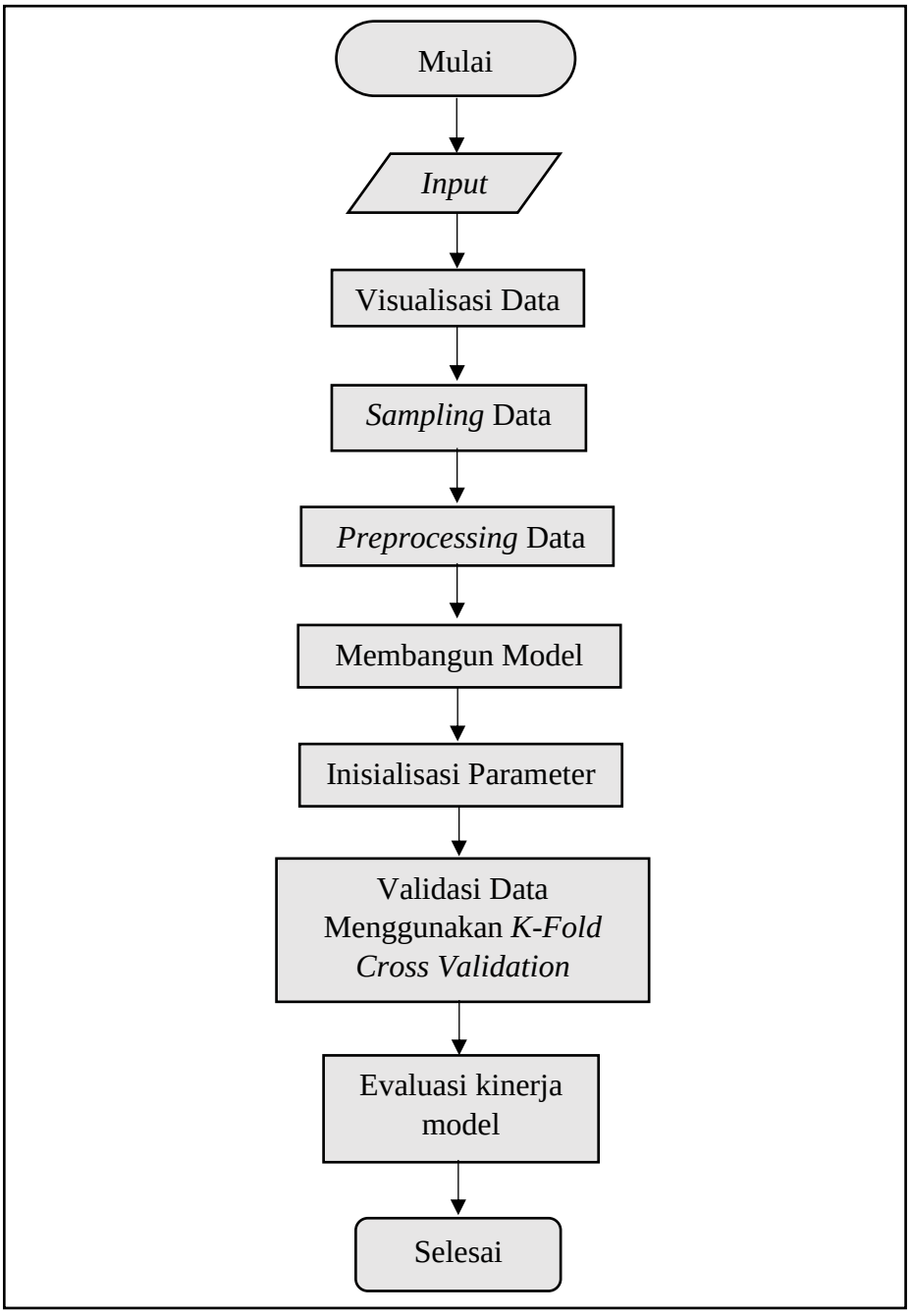
1. Melakukan *input* data judul berita berbahasa Indonesia mengenai COVID-19 yang sudah dilabeli untuk dilakukan klasifikasi dengan metode CNN dan LSTM.
2. Melakukan visualisasi data dengan membuat diagram batang untuk melihat perbandingan jumlah berita *event* dan berita *non-event* secara keseluruhan.
3. Melakukan pengambilan sampel acak sederhana (*simple random sampling*) sebagai simulasi untuk membentuk beberapa rasio ketidakseimbangan data. Rasio keseluruhan berita *event* terhadap berita *non-event* adalah 37%. Jumlah berita *event* akan dikurangi dengan cara mengambil sampel dari berita *event* sehingga akan terbentuk beberapa rasio berita *event* terhadap berita *non-event* yaitu, 30%, 20%, 10%, dan 1%.
4. Melakukan *preprocessing* data, yaitu *case folding*, *cleaning*, *stopword removal*, dan tokenisasi.
5. Melakukan *word embedding* menggunakan model *Bag of Words* (BoW) dengan algoritma *Count Vectorizer*.
6. Membangun model CNN dan LSTM untuk dilakukan *hyperparameter tuning.* Arsitektur model CNN dan LSTM yang digunakan pada penelitian ini diperlihatkan pada Gambar 7 dan Gambar 8.
7. Melakukan *hyperparameter tuning* agar mendapatkan kombinasi parameter terbaik untuk model yang sudah dibangun.
8. Melakukan validasi data menggunakan *K-Fold Cross-Validation*.
9. Mengevaluasi kinerja metode CNN dan LSTM dalam mengklasifikasi data judul berita mengenai COVID-19 berdasarkan akurasi, presisi, *recall,* dan *f1-score* untuk menemukan model yang paling andal dalam mengklasifikasi data teks dengan beberapa rasio ketidakseimbangan data.



Gambar 7. Arsitektur Model CNN.



Gambar 8. Arsitektur Model LSTM.



Gambar 9. Diagram Alir Metode Penelitian.

V. KESIMPULAN

5.1 Kesimpulan

Berdasarkan hasil dan pembahasan yang sudah dipaparkan pada bab sebelumnya, dapat diambil kesimpulan untuk penelitian ini sebagai berikut:

1. Klasifikasi teks judul berita COVID-19 menggunakan metode CNN dan LSTM dengan *word embedding layer* dan divalidasi dengan 10-*fold cross validation* memiliki kinerja yang baik dengan beberapa tingkat ketidakseimbangan data.
2. Kinerja dan waktu komputasi dari metode CNN mengungguli metode LSTM dalam mengklasifikasi data teks yang tidak seimbang pada tingkat ketidakseimbangan data 37%, 20%, dan 10%. Sehingga pada penelitian ini, metode CNN dapat diandalkan dalam mengklasifikasi data teks dengan beberapa tingkat ketidakseimbangan data.

5.2 Saran

1. Metode CNN merupakan metode yang dapat mengekstraksi fitur dari data teks, namun sering mengabaikan keterkaitan antar konteks pada teks. Metode LSTM dapat menjadi solusi karena memiliki kelebihan untuk mengenali keterkaitan antar konteks pada teks, tetapi tidak dapat mengekstraksi fitur dari data teks. Oleh karena itu, penulis menyarankan untuk mengombinasikan metode CNN dan LSTM pada penelitian klasifikasi teks selanjutnya agar memperoleh kelebihan dari metode CNN dan LSTM serta memiliki kinerja yang lebih baik.

DAFTAR PUSTAKA

Ahmed, J., dan Ahmed, M. 2021. Online News Classification Using Machine Learning Techniques. *IIUM Engineering Journal*. **22**(2): 210-225.

Asiyah, S. N. 2016. Klasifikasi Berita Online Menggunakan Metode Support Vector Machine Dan K-Nearest Neighbor. Disertasi. Institut Teknologi Sepuluh Nopember, Surabaya.

Brownlee, J. 2018. *Better deep learning: train faster, reduce overfitting, and make better predictions*. Machine Learning Mastery.

Cholissodin, I., Soebroto, A. A., Hasanah, U., dan Febiola, Y. I. 2020. *Ai, Machine Learning & Deep Learning*.

Fadli, K., dan Novita, P. 2021. Analisis Framing Media Online Tentang Pandemi Covid-19 (Studi Kasus Covid-19 Pada Media Online Tribun News. com dan Kepri. co. id Edisi Bulan Maret s/d Juni 2020). *Jurnal Purnama Berazam*. **2**(2): 172-200.

Faturohman, F., Irawan, B., dan Setianingsih, C. 2020. Analisis Sentimen Pada Bpjs Kesehatan Menggunakan Recurrent Neural Network. *eProceedings of Engineering*. **7**(2).

Hairani, Setiawan, N. A., dan Adji, T. B. 2019. Metode Klasifikasi Data Mining dan Teknik Sampling SMOTE Menangani Class Imbalance Untuk Segmentasi Customer Pada Industri Perbankan. in Proceedings SNST Fakultas Teknik.

Johnson, J. M., dan Khoshgoftaar, T. M. 2019. Survey on deep learning with class imbalance. *Journal of Big Data*. **6**(1): 1–54.

Kapil, P., Ekbal, A., dan Das, D. 2020. Investigating deep learning approaches for hate speech detection in social media. *arXiv preprint arXiv:2005.14690*.

Kemenkes RI. 2020. *Keputusan menteri kesehatan republik indonesia nomor hk. 01.07/menkes/413/2020 tentang pedoman pencegahan dan pengendalian corona virus disease 2019 (covid-19).* Menteri Kesehatan Republik Indonesia.

Khomsah, S., dan Aribowo, A. S. 2020. Text-preprocessing model youtube comments in Indonesian. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*. **4**(4): 648-654.

Khotimah, P. H., Rozie, A. F., Nugraheni, E., Arisal, A., Suwarningsih, W., dan Purwarianti, A. 2020. Deep Learning for Dengue Fever Event Detection Using Online News, hlm 261-266. In *2020 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET)*.

Kurniawan, T. 2017. Implementasi Text Mining Pada Analisis Sentimen Pengguna Twitter Terhadap Media Mainstream Menggunakan Naïve Bayes Classifier Dan Support Vector Machine. Disertasi. Institut Teknologi Sepuluh Nopember, Surabaya.

L'heureux, A., Grolinger, K., Elyamany, H. F., dan Capretz, M. A. 2017. Machine learning with big data: Challenges and approaches. *IEEE Access*. **5**: 7776-7797.

Launa, L. 2020. Banjir Infodemi: Viralitas Akurasi Berita Virologi Dalam Fenomena Coronavirus Disease. *The Source: Jurnal Ilmu Komunikasi*. **2**(2): 1-21.

Levy, P. S., dan Lemeshow, S. 2008. *The population and the sample, Sampling of Populations: Methods and applications*. 4th ed. John Wiley and Sons, New York, USA.

Liu, H., Zhou, M., dan Liu, Q. 2019. An embedded feature selection method for imbalanced data classification. *IEEE/CAA Journal of Automatica Sinica*. **6**(3): 703-715.

Minaee, S., Kalchbrenner, Cambria, N., E., Nikzad, N., Chenaghlu, M., dan Gao, J. 2021. Deep learning–based text classification: a comprehensive review. *ACM Computing Surveys (CSUR)*. **54**(3): 1–40.

Nugraheni, E., Khotimah, P. H., Arisal, A., Rozie, A. F., Riswantini, D., dan Purwarianti, A. 2020. Classifying aggravation status of COVID-19 event from short-text using CNN, hlm. 240-245. In 2020 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET).

Nurdin, A., Aji, B. A. S., Bustamin, A., dan Abidin, Z. 2020. Perbandingan Kinerja Word Embedding Word2Vec, Glove, Dan Fasttext Pada Klasifikasi Teks. *Jurnal Tekno Kompak*. **14**(2): 74-79.

Olah, C. 2015. Understanding LSTM Networks, Github, 27 August 2015. https://colah.github.io/posts/2015-08-Understanding-LSTMs/. Diakses pada 15 Maret 2022.

Parolo, D. 2020. *Deep learning for text classification: an application of generalized language models for italian texts*.

Rachman, F. P. 2021. Perbandingan Model Deep Learning Untuk Klasifikasi Sentiment Analysis Dengan Teknik Natural Languange Processing. *Jurnal Teknologi dan Manajemen Informatika*. **7**(2).

Ramadhanti, N. S., Kusuma, W. A., dan Annisa, A. 2020. Optimasi Data Tidak Seimbang pada Interaksi Drug Target dengan Sampling dan Ensemble Support Vector Machine. *Jurnal Teknologi Informasi dan Ilmu Komputer*. **7**(6): 1221-1230.

Rozi, I. F., Wijayaningrum, V. N., dan Khozin, N. 2020. Klasifikasi Teks Laporan Masyarakat Pada Situs Lapor! Menggunakan Recurrent Neural Network. *Sistemasi: Jurnal Sistem Informasi*. **9**(3): 633-645.

Santos, M. S., Soares, J. P., Abreu, P. H., Araujo, H., dan Santos, J. 2018. Cross-validation for imbalanced datasets: avoiding overoptimistic and overfitting approaches [research frontier]. *IEEE Computational Intelligence Magazine*. **13**(4): 59-76.

Shaikh, S., Daudpota, S. M., Imran, A. S., dan Kastrati, Z. 2021. Towards improved classification accuracy on highly imbalanced text dataset using deep neural language models. *Applied Sciences*. **11**(2): 869.

Sharma, S., Sharma, S., dan Athaiya, A. 2017. Activation functions in neural networks. *Towards data science*. **6**(12): 310-316.

Susanty, M., dan Sukardi, S. 2021. Perbandingan Pre-trained Word Embedding dan Embedding Layer untuk Named-Entity Recognition Bahasa Indonesia. *Petir*. **14**(2): 247-257.

Taeho, J. 2019. Text mining concepts, implementation, and big data challange,(p. 1). Hongik University, Seoul, Korea.

Thompson, *S. K. Sampling/by steven k. thompson*. Tech. Rep.

Wahyudi, M. D. R., Fatwanto, A., Kiftiyani, U., dan Wonoseto, M. G. 2021. Topic Modeling of Online Media News Titles during COVID-19 Emergency Response in Indonesia Using the Latent Dirichlet Allocation (LDA) Algorithm. *Telematika.* **14**(2): 101-111.

Wang, Q., Li, W., dan Jin, Z. 2021. Review of Text Classification in Deep Learning. *Open Access Library Journal*. **8**(3): 1-8.

Widhiyasana, Y., Semiawan, T., Mudzakir, I. G. A., dan Noor, M. R. 2021. Penerapan Convolutional Long Short-Term Memory untuk Klasifikasi Teks Berita Bahasa Indonesia. *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*. **10**(4): 354-361.

Yechuri, P. K., dan Ramadass, S. 2021. Classification of Image and Text Data Using Deep Learning-Based LSTM Model. *Traitement du Signal*. **38**(6).