GENERATED TEXT DALAM RESTRUKTURISASI KETIDAKSEIMBANGAN DATASET JUDUL BERITA *COVID-19* DENGAN MENGGUNAKAN MODEL *DEEP LEARNING*

(Skripsi)

Oleh

FEBI SITI SUTRIA NINGSIH
NPM 1817031078

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2022

ABSTRACT

GENERATED TEXT IN RESTRUCTURING AN IMBALANCED DATASET TITLE COVID-19 NEWS USING DEEP LEARNING MODEL

By

FEBI SITI SUTRIA NINGSIH

Imbalanced data is a problem that often occurs in classifying. This has the potential to cause bias against the majority class. Oversampling or text generation is one of the methods used to overcome imbalanced data. This study proposes the creation of a synonym-based text to restructure the imbalanced covid-19 online news dataset. Deep learning models (CNN and LSTM) use the word embedding feature and validated using the k-fold cross validation method for imbalanced and balanced data. The results showed that using both imbalanced and balanced data, CNN achieved significantly higher performance by 7% - 9% in accuracy, precision, recall, and f1-score than LSTM. Although the text generation method is proven to be able to overcome the problem of data imbalance, the text generation method has a drawback, namely that there is often a loss of semantic meaning.

Keywords: Text Generation, Imbalanced Dataset, Deep Learning, k-Fold Cross Validation.

ABSTRAK

GENERATED TEXT DALAM RESTRUKTURISASI KETIDAKSEIMBANGAN DATASET JUDUL BERITA *COVID-19* DENGAN MENGGUNAKAN MODEL *DEEP LEARNING*

Oleh

FEBI SITI SUTRIA NINGSIH

Imbalanced data merupakan suatu permasalahan yang sering terjadi dalam melakukan klasifikasi. Hal ini berpotensi menyebabkan bias terhadap kelas mayoritas. *Oversampling* atau *text generation* adalah salah satu metode yang digunakan untuk mengatasi *imbalanced* data. Penelitian ini mengusulkan pembuatan teks berbasis sinonim untuk merestrukturisasi dataset berita *online covid-19* yang tidak seimbang. Model *deep learning* (CNN dan LSTM) yang menggunakan fitur *word embedding* dan divalidasi menggunakan metode *k-fold cross validation* untuk *imbalanced* dan *balanced* data. Hasil penelitian menunjukkan bahwa baik dengan menggunakan data *imbalanced* dan *balanced* data, CNN mencapai kinerja yang secara signifikan lebih tinggi sebesar 7% - 9% dalam akurasi, presisi, *recall*, dan *f1-score* daripada LSTM. Meskipun metode *text generation* terbukti dapat mengatasi masalah ketidakseimbangan data, tetapi metode *text generation* memiliki kekurangan yaitu sering terjadi kehilangan makna semantik.

Keywords: *Text Generation*, Ketidakseimbangan Data, *Deep Learning*, *k-Fold Cross Validation.*

GENERATED TEXT DALAM RESTRUKTURISASI KETIDAKSEIMBANGAN DATASET JUDUL BERITA *COVID-19* DENGAN MENGGUNAKAN MODEL *DEEP LEARNING*

Oleh

FEBI SITI SUTRIA NINGSIH

Skripsi

Sebagai Salah Satu Syarat untuk Mencapai Gelar
SARJANA MATEMATIKA

Pada

Jurusan Matematika
Fakultas Matematika Dan Ilmu Pengetahuan Alam
Universitas Lampung

JURUSAN MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
2022

Judul Skripsi	:	***GENERATED TEXT* DALAM RESTRUKTURISASI KETIDAKSEIMBANGAN DATASET JUDUL BERITA *COVID-19* DENGAN MENGGUNAKAN MODEL *DEEP LEARNING***
Nama Mahasiswa	:	**FEBI SITI SUTRIA NINGSIH**
Nomor Pokok Mahasiswa	:	**1817031078**
Program Studi	:	**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 Oktober 2022**

PERNYATAAN SKRIPSI MAHASISWA

Saya yang bertanda tangan dibawah ini:

Nama	: Febi Siti Sutria Ningsih
Nomor Pokok Mahasiswa	: 1817031078
Jurusan	: Matematika
Judul Skripsi	: ***Generated Text* Dalam Restrukturisasi Ketidakseimbangan Dataset Judul Berita *Covid-19* dengan Menggunakan Model *Deep Learning***

Dengan ini menyatakan bahwa skripsi ini adalah hasil pekerjaan saya sendiri, bukan hasil orang lain. Apabila dikemudian 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, 2 Desember 2022

Penulis

Febi Siti Sutria Ningsih
NPM. 1817031078

RIWAYAT HIDUP

Penulis bernama Febi Siti Sutria Ningsih, dilahirkan di Kota Jakarta, Provinsi Daerah Khusus Ibukota Jakarta pada 12 Februari 2001. Penulis merupakan anak kedua dari pasangan Bapak Sutriyono dan Ibu Soleha Suhartinah.

Penulis mengawali pendidikan di Taman Kanak-Kanak (TK) Islam Al-Wakaf pada tahun 2005-2006. Kemudian menempuh pendidikan Sekolah Dasar (SD) di SD Negeri 11 Pagi di Jakarta Pusat pada tahun 2006-2008, lalu melanjutkan pendidikan di SD Negeri 05 Pagi Jakarta Barat pada tahun 2008-2011, dan terakhir melanjutkan pendidikan di SD Negeri Kadudampit 1 di Pandeglang pada tahun 2011-2012. Melanjutkan ke Sekolah Menengah Pertama (SMP) di SMP Negeri 1 Saketi pada tahun 2012-2013 dan pindah ke SMP Negeri 1 Blambangan Umpu lulus pada tahun 2015. Sekolah Menengah Atas di SMAN 1 Blambangan Umpu lulus pada tahun 2018.

Pada tahun 2018 penulis terdaftar sebagai mahasiswa S1 Jurusan Matematika FMIPA UNILA melalui jalur Penerimaan Mahasiswa Perluasan Akses Pendidikan (PMPAP). Selama menjadi mahasiswa penulis juga aktif dalam organisasi Pers Mahasiswa Natural FMIPA UNILA. Pada tahun 2021 penulis melakukan Kuliah Praktik di Badan Pusat Statistik (BPS) Kabupaten Way Kanan dan Kuliah Kerja Nyata (KKN) di Desa Lembasung, Kecamatan Blambangan Umpu, Kabupaten Way Kanan.

KATA INSPIRASI

"Dan bersabarlah, Sesungguhnya Allah beserta orang-orang yang sabar."
(Q.S. Al Anfaal: 46)

"Rasulullah bersabda: Barangsiapa menempuh jalan untuk mendapatkan ilmu, Allah akan memudahkan baginya jalan menuju surga."
(HR. Muslim)

"The problems and the worries that you created in your head, they're all delutions."
(*Mark Lee*)

"It's not easy but that's a life. Be strong cause there are better days ahead. Lets be grateful for what we have."
(*Mark Lee*)

"You gotta seize the opportunity."
(Penulis)

PERSEMBAHAN

Dengan mengucapkan rasa syukur atas segala puji dan kehadirat Allah SWT. yang telah memberikan rahmat dan hidayah-Nya sehingga penulis dapat menyelesaikan skripsi ini. Serta tak lupa shalawat serta salam selalu tercurahkan kepada junjungan kita Nabi Muhammad SAW.

Dengan penuh ketulusan, penulis mempersembahkan skripsi ini untuk:

Ibu Soleha Suhartinah dan Ayah Sutriyono

Orang tua yang selalu memberikan dukungan dalam bentuk apapun dalam setiap pengambilan keputusan dan keadaan febi serta doa yang selalu dipanjatkan.

Kakak Fajar Sholehudin Al-Ayubi dan Adik Firdatun Hasanah Prasetia

Kakak dan adik yang seperti teman selama dirumah, pemberi motivasi yang baik disetiap keadaan.

Dosen Pembimbing dan Pembahas

Terima kasih kepada dosen pembimbing dan pembahas yang selalu sabar dan membantu, memberikan arahan, masukkan, dan ilmu yang bermanfaat.

Teman-Teman Jurusan Matematika Angkatan 2018

Almamater Tercinta Universitas Lampung

SANWACANA

Puji syukur kehadirat Allah SWT. atas rahmat dan hidayah-Nya, shalawat serta salam selalu tercurahkan kepada junjungan kita Nabi Muhammad SAW. sehingga penulis dapat menyelesaikan skripsi dengan judul **"*Generated Text* Dalam Restrukturisasi Ketidakseimbangan Dataset Judul Berita *Covid-19* Dengan Menggunakan Model *Deep Learning*"**. Penulis menyadari bahwa skripsi ini tidak akan terselesaikan tanpa adanya bantuan, bimbingan, serta saran dari berbagai pihak. Oleh karena itu, penulis ingin menyampaikan banyak terima kasih kepada:

1. Ibu Dian Kurniasari, S.Si., M.Sc. selaku dosen pembimbing I dan pembimbing akademik yang selalu memberikan bantuan, motivasi, masukan dan saran yang mendukung sehingga penulis dapat menyelesaikan skripsi ini.
2. Ibu Dr. Purnomo Husnul Khotimah, M.T. selaku dosen pembimbing II atas saran dan masukkan yang membantu penulis menyelesaikan skripsi ini.
3. Bapak Ir. Warsono, M.S., Ph.D. selaku dosen penguji yang telah memberikan kritik dan saran yang membangun selama proses penyusunan skripsi.
4. Bapak Dr. Aang Nuryaman, S.Si., M.Si. selaku Ketua Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.
5. Bapak Dr. Eng. Suripto Dwi Yuwono, S.Si., M.T. selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.
6. Seluruh dosen, staff, karyawan Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.
7. Seluruh staff keltian *Information Retrieval* (IR), unit kerja Pusat Riset Informatika, Badan Riset dan Inovasi Nasional (BRIN).
8. Ibu dan Ayah yang tiada hentinya memberikan doa, dukungan dan motivasi kepada penulis.
9. Kakak dan adikku yang turut memberikan dukungan dan semangat kepada penulis.

10. Teman-teman seperbimbingan skripsi yaitu Dalfa, Farrel, Ferzy, Joshua, Luthfia, Maydia, Nur Alifiah, Oktina, Putri Salsabila, Rekti, Sulistian, Virda, Zaenal yang memberikan bantuan dan masukkan kepada penulis.
11. Sahabat-sahabatku selama masa perkuliahan yaitu Aniisah, Ahya, Dewi Kartika, Eva, Dewi Utari, Dora Panny, Hilda, Nurul dan yunda Mayda yang menemani dan membersamai penulis selama masa perkuliahan.
12. Semua teman jurusan Matematika angkatan 2018 dan teman kelas B yang tidak bisa disebutkan satu persatu telah memberikan semangat kepada penulis.
13. Tria Yuliani, Windy Devi Yantri dan Lily Rahmawati sebagai sahabat penulis sejak SMP yang selalu menemani di masa sulit dan memberikan doa serta dukungan.
14. Seluruh pihak terkait yang telah banyak membantu dan tidak dapat disebutkan satu persatu.
15. *Last but not least* terima kasih kepada diri sendiri karena telah bertahan hingga akhir mendapatkan gelar sarjana. Segala suka dan duka dalam menjalani masa perkuliahan akan selalu diingat dan dijadikan pelajaran untuk tantangan selanjutnya.

Penulis menyadari dalam penyusunan skripsi ini masih terdapat banyak kekurangan. Oleh karena itu, kritik dan saran sangat diharapkan untuk menyempurnakan skripsi ini.

Bandar Lampung, November 2022
Penulis,

Febi Siti Sutria Ningsih

DAFTAR ISI

Halaman

DAFTAR TABEL

Tabel Halaman

DAFTAR GAMBAR

Gambar Halaman

I. PENDAHULUAN

1.1 Latar Belakang dan Masalah

Machine learning merupakan teknik inferensi data yang menggunakan pendekatan matematis untuk membuat model yang mencerminkan pola data. *Machine learning* minimum memiliki dua tujuan yaitu untuk memprediksi masa depan dan memperoleh pengetahuan (Putra, 2018). Terdapat beberapa masalah dalam pemrosesan data pada *machine learning* salah satunya yaitu ketidakseimbangan kelas data.

Ketidakseimbangan kelas data terjadi karena terdapat kelas yang berisi sejumlah besar titik data (mayoritas) dan kelas lainnya minoritas. Jika rasio data tidak seimbang, maka akan mengalami kesulitan untuk mengklasifikasikan data dalam kelas minoritas. Karena untuk meminimalkan tingkat kesalahan algoritma pembelajaran hanya akan mengandalkan kelas mayoritas dalam pengklasifikasian (Mutmainah, 2021). Pada data yang tidak seimbang, distribusi setiap kelas tidak merata karena beberapa kelas lebih sering muncul dibandingkan dengan kelas lainnya. Kondisi ini menyebabkan *classifier* menunjukkan bias terhadap kelas mayoritas dan dalam beberapa kasus ekstrim dapat mengabaikan kelas minoritas sama sekali. Perilaku *classifier* pada kelas tidak seimbang tidak menguntungkan karena seringkali data kelas minoritas adalah kelas yang memiliki informasi penting (Shaikh, dkk., 2021).

Oleh karena itu, kondisi kelas tidak seimbang perlu diatasi sebelum melakukan klasifikasi. Mengatasi ketidakseimbangan kelas, ada beberapa pendekatan yang dapat dilakukan salah satunya adalah pendekatan data dengan cara mengurangi (*undersampling*) atau menambahkan data (*oversampling*). Jika melakukan

undersampling, beberapa data pada kelas mayoritas akan dihilangkan sehingga terdapat kemungkinan hilangnya data berharga pada kelas mayoritas. Sedangkan teknik *oversampling* dilakukan dengan pendistribusian data yang seimbang dengan cara replikasi data minoritas secara acak. Akan tetapi, kekurangan dari melakukan *oversampling* adalah dapat memungkinkan munculnya *overfitting* (Akbar dkk., 2019).

Secara umum, metode *oversampling* memberikan hasil klasifikasi yang lebih baik daripada metode *undersampling* (Siringoringo, 2018). Chawla dkk (2002) mengajukan solusi menangani *overfitting* pada metode *oversampling* yaitu menggunakan metode *Synthetic Minority Oversampling Technique* (SMOTE) dengan memanfaatkan *nearest neighbors* serta jumlah *oversampling* untuk pendekatan data numerik. Pada tahun 2020, Arifiyanti dan Wahyuni melakukan penelitian dengan dataset yang tidak seimbang menggunakan SMOTE dan model klasifikasi *logistic linear*, KNN, *naive bayes* dan *decision tree*. Hasil dari penelitian ini yaitu model klasifikasi yang dihasilkan menunjukkan bahwa metode SMOTE meningkatkan performa model klasifikasi. Sedangkan pada tahun 2021, Sulistiyono dkk melakukan penelitian mengenai data yang tidak seimbang dengan menggunakan SMOTE menghasilkan nilai akurasi dan *g-mean* yang paling baik.

Teknik untuk menghasilkan data bangkitan dan *oversampling* seperti SMOTE dapat mengatasi masalah ketidakseimbangan data, tetapi metode ini memiliki kekurangan yaitu terdapat *overfitting* dan *noise* yang signifikan. Metode ini telah terbukti dapat menghasilkan data numerik dan gambar buatan, tetapi efektivitas yang dibutuhkan untuk data teks agar dapat mempertahankan struktur tata bahasa, konteks, dan informasi semantik belum dievaluasi kembali (Shaikh dkk., 2021). Selain itu, dalam hal bahasa Indonesia masih terdapat masalah kelangkaan sumber daya yaitu terbatasnya ketersediaan koleksi data dan perpustakaan *Natural Language Processing* (NLP) (Wilie dkk,. 2020). Kemudian untuk validasi membutuhkan penggunaan *k-fold cross validation*. Hal ini dikarenakan untuk mengetahui performa model dengan melakukan percobaan sebanyak *k* kali yang berarti data dibagi ke dalam *k* bagian dengan komposisi kelas yang seimbang

setiap bagiannya. Hal ini dapat membantu penelitian karena *cross validation* mampu bekerja dengan cepat sehingga jumlah pengujian data latih dan data uji akan diambil dengan percobaan/iterasi sebelumnya (Santosa dan Umam, 2018).

Berdasarkan uraian diatas, penelitian ini akan melakukan *oversampling* dengan metode pembuatan teks menggunakan penggantian sinonim *n*-kata kemudian hasil data yang telah seimbang dan tidak seimbang akan divalidasi dengan menggunakan *k-fold cross validation*. Selain itu, penelitian ini juga akan melakukan uji evaluasi akurasi, presisi, *recall*, dan *f1-score* dengan model *deep learning* LSTM dan CNN terhadap data yang telah dilakukan *resampling*. Hasil uji evaluasi klasifikasi tersebut akan dibandingkan dengan hasil uji klasifikasi tanpa *resampling*.

1.2 Tujuan Penelitian

Adapun tujuan dari penelitian ini diantaranya:

1. Mengetahui performa *text generation* untuk menyeimbangkan kelas pada dataset menggunakan model *deep learning* dan validasi dengan *k-fold cross validation.*
2. Membandingkan uji evaluasi model klasifikasi *deep learning* antara data *imbalanced* dengan data *balanced* setelah divalidasi menggunakan *k-fold cross validation*.

1.3 Manfaat Penelitian

Adapun manfaat dari penelitian ini yaitu:

1. Sebagai rujukan pengembangan ilmu matematika dalam mengatasi permasalahan *imbalanced dataset* dengan menggunakan *text generation.*
2. Memberikan pengetahuan pada bidang ilmu data mining khususnya tentang pengaplikasian *text mining* dengan model *deep learning.*

II. TINJAUAN PUSTAKA

2.1 *Text Mining*

Text mining yaitu salah satu teknik penggalian informasi untuk menemukan pola yang menarik dari sekumpulan data berupa teks dalam bentuk klasifikasi maupun *clustering*. *Text mining* masih merupakan bagian dari data mining dimana akan memproses data teks serta dokumen-dokumen yang bisa jadi dalam jumlah yang sangat besar (Harjanta, 2015). Bentuk dari data yang dimasukkan pada *text mining* merupakan data-data yang tidak terstruktur, seperti dokumen XML, PDF, dan sejenisnya (Adhi dkk., 2019).

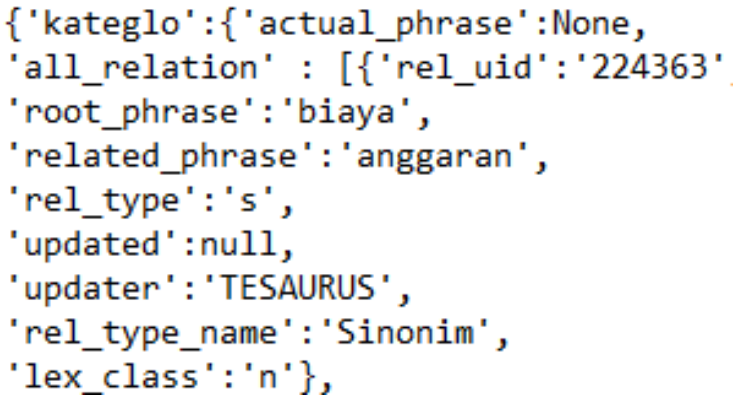
2.2 Ketidakseimbangan Data

Ketidakseimbangan kelas data terjadi ketika rasio dalam sebuah kelas besar dan kelas lainnya lebih kecil. Penggunaan data yang tidak seimbang akan mengakibatkan kelas minoritas menghasilkan nilai akurasi yang rendah (Ardiyansyah dan Rahayuningsih, 2020). Tingkat kesalahan dapat diminimalkan dengan cara mengklasifikasikan semua contoh ke dalam kelas mayoritas tetapi kelas minoritas akan diidentifikasi sebagai klasifikasi yang salah (Akbar dkk., 2019). Permasalahan ketidakseimbangan kelas data dapat diselesaikan dengan 3 pendekatan untuk menyelesaikan permasalahan ketidakseimbangan kelas data yaitu pendekatan data, pendekatan algoritma, dan penggabungan antara pendekatan data dan algoritma. Pendekatan teknik data yaitu dengan cara menyeimbangkan distribusi data baik menggunakan *oversampling* (membangkitkan data sintetis) maupun *undersampling* (mengurangi jumlah data pada kelas mayoritas). Pendekatan algoritma yaitu dengan mengembangkan

algoritma baru atau memodifikasi algoritma yang telah ada sebelumnya (Ardiyansyah dan Rahayuningsih, 2020).

2.3 Kateglo API

Kateglo API adalah layanan web yang diakses menggunakan *Application Programming Interface* (API) menyediakan definisi, sinonim, antonim, dan glosarium yang berkaitan dengan kata berbasis data PHP dan MySQL (Oyong dkk., 2018). Nama kateglo berasal dari singkatan kamus, tesaurus, dan glosarium. Kateglo terdiri dari 72.253 entri kamus, 191.000 entri glosarium, 2.012 entri peribahasa, dan beberapa entri kamus bahasa Indonesia dengan 3.423 entri singkatan dan akronim. Kateglo dapat digunakan dengan memasukkan kata kunci yang diperlukan melalui URL: http://kateglo.com/api.php?format=json& phrase=[word] dan menyediakan *output* dalam format JSON atau XML (Rohman dkk., 2019). Gambar 1 merupakan contoh *output* kateglo jika kata yang dicari adalah biaya.



Gambar 1. Contoh keluaran Kateglo API

2.4 *Word Embedding*

Word embedding merupakan teknik pembelajaran fitur dalam NLP untuk membangun representasi vektor kata dari kumpulan teks. *Word embdding* menawarkan representasi yang lebih ekspresif dan efisien dengan mempertahankan kesamaan konteks kata dengan membangun vektor berdimensi rendah (Naili dkk., 2017). Proses pembelajaran representasi vektor kata dapat disisipkan saat pembentukan model. Beberapa contoh *word embedding* yaitu *embedding layer, word2vec,* dan *glove* (Gelar dan Sari., 2020). *Word embedding* yang digunakan pada penelitian ini yaitu *Bag of Words* (BoW).

Bag of Words merupakan salah satu metode paling sederhana dalam mengubah data teks menjadi vektor yang dapat dipahami oleh komputer. Metode ini sejatinya hanya menghitung frekuensi kemunculan kata pada seluruh dokumen. Karena dalam BoW kata yang terlihat tidak perlu diolah akan dihilangkan. Sehingga saat waktu pemrosesan tidak memerlukan waktu yang lama (Guritno dan Santosa, 2017).

2.5 *Deep Learning*

Deep learning merupakan salah satu teknik dalam *machine learning* yang mencoba memodelkan data dengan arsitektur lebih mendalam dibanding dengan teknik *machine learning* lainnya dalam menyelesaikan masalah prediksi maupun klasifikasi (Hao dkk, 2016). *Deep neural network* merupakan bentuk dari jaringan sistem syaraf tiruan yang memiliki 3 lapisan atau lebih. Oleh karena itu, sulit dilakukan estimasi parameter karena jaringan ini memiliki banyak lapisan.

Deep learning dapat menyelesaikan permasalahan yang tidak dapat diselesaikan *multilayer perceptron* yaitu menentukan relasi tersembunyi antara *input* dan

output. Proses pembelajaran pada *deep learning* cenderung lebih lama karena memiliki banyak parameter. Teknik yang sering digunakan adalah *successive learning* yaitu membangun suatu jaringan secara bertahap seperti saat melatih jaringan neural menggunakan 3 lapisan, lalu ditambahkan 1 lapisan lagi sehingga menjadi 4 lapisan dan seterusnya (Putra, 2018).

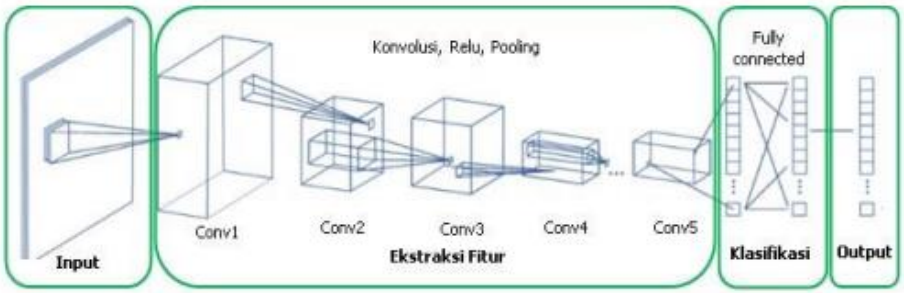
2.6 Inisialisasi Parameter (*Hypertunning*)

Hypertunning perlu dilakukan sebelum melakukan pembelajaran pada model. Penyesuaian *hyperparameter* model berguna untuk memaksimalkan kinerja model saat akan melakukan proses validasi. Saat menjalankan proses *hypertunning*, memerlukan adanya perintah *early stopping* agar ketika parameter sudah mencapai nilai yang optimal proses iterasi akan berhenti. Proses *hypertunning* yang dilakukan seperti menambahkan *dropout*, *epochs*, *batch size* dan menentukan jumlah neuron. Berikut merupakan parameter yang digunakan pada proses *hypertunning* yaitu:

a. *Dropout* digunakan untuk mencegah adanya *overfitting* pada model dan mempersingkat waktu saat proses pembelajaran. Sistem kerja dropout dengan menghilangkan sementara suatu neuron yang merupakan *hidden layer* pada *neural* (Minarno dkk., 2021).
b. *Epoch* dapat disebut sebagai iterasi yang dimana ketika seluruh dataset telah melalui proses *training* pada model dan dikembalikan ke awal putaran pertama (Brownlee, 2016).
c. *Batch size* merupakan pembagian *epoch* dalam bagian-bagian kecil untuk mempercepat proses *hypertunning*. Hal ini dilakukan karena untuk memproses satu *epochs* membutuhkan waktu yang lama. Oleh karena itu diperlukan adanya *batch size* agar saat proses *training* dilakukan pembagian per *batch* (Nugroho, dkk., 2020).

2.7 *Convolutional Neural Network* (CNN)

Convolutional Neural Network merupakan salah satu penerapan dari model *deep learning* dengan jaringan saraf yang tersusun dari 3 lapisan yaitu *input layer*, *output layer*, dan *hidden layer* (Suartika dkk., 2016). Di dalam *hidden layer* terdapat lapisan yang tersusun secara bertumpuk yaitu *convolutional layer*, *pooling layer*, dan *fully connected layer* (Suyanto, 2018). Kontribusi utama dari CNN adalah pada lapisan *convolutional* dan lapisan *pooling* (Suartika dkk., 2016).



Gambar 2. Arsitektur CNN (Sumber: Krizhevsky dkk, 2012)

a. *Convolution Layer*

Convolutional Neural Network menggunakan filter yang disebut sebagai kernel untuk mendeteksi fitur-fitur. Filter hanyalah sebuah matriks nilai yang disebut bobot, dilatih untuk mendeteksi fitur tertentu. Untuk memberikan nilai yang menunjukkan seberapa tepat fitur yang tersedia tersebut, filter melakukan operasi konvolusi (Kadir dan Susanto, 2013).

b. *Pooling Layer*

Pooling layer berada setelah lapisan konvolusional yang digunakan untuk merangkum informasi yang dihasilkan dari lapisan konvolusional. Vektor yang dihasilkan akan dikombinasikan dan menjadi vektor baru. Tujuan dari penggunaan *pooling layer* untuk mengurangi jumlah parameter karena lapisan *pooling* yang dimasukkan diantara lapisan konvolusi secara berturut-turut dalam arsitektur model CNN dapat mengurangi ukuran volume *output* pada *feature map* perhitungan jaringan untuk mengendalikan *overfitting* (Suartika dkk., 2016).

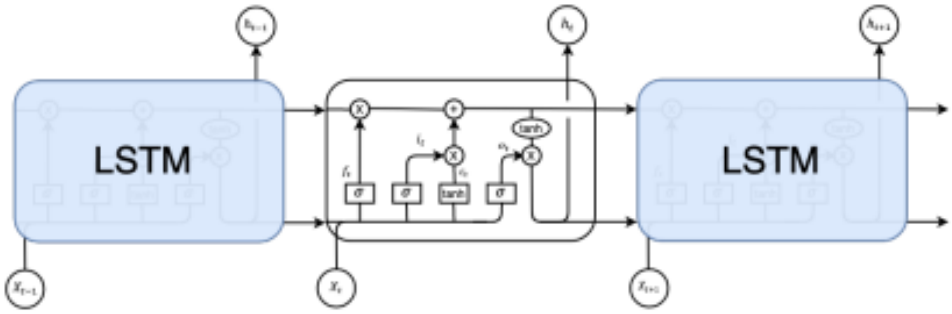
c. *Fully-Connected Layer*

Fully-Conected Layer adalah sebuah lapisan dimana kumpulan dari proses konvolusi. Perbedaan antara lapisan *fully connected* dan lapisan konvolusi biasa adalah neuron di lapisan konvolusi terhubung hanya ke daerah tertentu pada *input*, sementara lapisan *fully connected* memiliki neuron yang secara keseluruhan terhubung. Namun, kedua lapisan tersebut masih mengoperasikan produk dot, sehingga fungsinya tidak begitu berbeda (Hijazi dkk., 2015). Fungsi dari lapisan ini yaitu untuk menyatukan semua *node* menjadi satu dimensi (Albewi dan Mahmood, 2017).

2.8 *Long Short Term Memory* (LSTM)

Menurut Hochreiter dan Schmidhuber (1997) LSTM adalah arsitektur jaringan saraf berulang khusus dengan kemampuan untuk mempelajari ketergantungan di antara data dalam kumpulan data. Poin utama LSTM adalah menyimpan informasi baru serta menjaga informasi lama agar tidak menghilang saat proses pengolahan data. Kelebihan dari LSTM yaitu menyimpan informasi jangka panjang dimana dalam setiap sel LSTM memiliki 3 gerbang pengatur yaitu *input*, *forget*, dan *output*. Gerbang *input* bertugas mengontrol masuk dan proses data dari luar, gerbang *forget* bertugas memutuskan data yang akan dihilangkan untuk proses selanjutnya, dan gerbang *output* melakukan seluruh perhitungan masuk kemudian menghasilkan keluaran dalam setiap sel LSTM (Zahara dan Sugianto, 2021).

Long Short Term Memory satu komponen gerbang digunakan saat mengontrol informasi yang masuk kedalam memori yang bertugas memecahkan masalah. Koneksi yang berulang menambah keadaan atau memori ke jaringan dan memungkinkanya untuk memanfaatkan pengamatan yang terurut (Hermanto dkk., 2021). Arsitektur dari LSTM terdiri dari lapisan *input*, lapisan *output*, dan lapisan tersembunyi seperti pada gambar berikut.



Gambar 3. Arsitektur LSTM (Sumber: Wiranda dan Sadikin, 2019)

Lapisan tersembunyi terdiri dari sel memori, satu sel memori memiliki tiga gerbang yaitu *input gate*, *forget gate*, dan *output gate* (Vinayakumar dkk., 2017). Gerbang *input* berfungsi mengontrol berapa banyak informasi yang harus disimpan dalam keadaan sel. Ini mencegah sel dari menyimpan data yang tidak perlu. Gerbang *forget* berfungsi mengontrol sejauh mana nilai tetap di dalam sel memori. Gerbang *output* berfungsi untuk memutuskan berapa banyak konten atau nilai dalam sel memori, digunakan untuk menghitung *output* (Mathisen, 2018).

a. *Input gate*

Input gate memiliki fungsi mengambil *output* sebelumnya dan kemudian dijadikan input baru melalui lapisan sigmoid. Rumus dari *input gate* yaitu (Wiranda dan Sadikin, 2019):

$$i_t = \sigma(W_i \times [h_{t-1}, x_t] + b_i) \quad (2.1)$$

dengan,

σ = Fungsi aktivasi sigmoid

Wi = Bobot dari *input iate*.

h_{t-1} = *State* sebelumnya atau *state* pada waktu t-1.

x_t = *Input* pada waktu t.

b_i = Bias pada *input gate*

kemudian nilai gerbang *input* dikalikan dengan *output* dari lapisan kandidat ($\tilde{C}$).

$$\tilde{C}_t = tanh(W_c \times [h_{t-1}, x_t] + b_c) \quad (2.2)$$

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \quad (2.3)$$

dengan,

$\tilde{C}$ = *Intermediate cell state*

Wc = Bobot dari *cell state*

h_{t-1} = *State* sebelumnya atau *state* pada waktu t – 1

x_t = *Input* pada waktu t

b. *Forget gate*

Forget gate adalah lapisan sigmoid yang mengambil *output* pada waktu t – 1 dan *input* pada waktu t dan menggabungkannya serta menerapkan fungsi aktivasi sigmoid. Karena sigmoid, *output* dari gerbang ini adalah 0 atau 1. Jika $f_t = 0$ maka keadaan (*state*) sebelumnya akan dilupakan, sementara jika $f_t = 1$ *state* sebelumnya tidak berubah. Rumus dari f_t adalah

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

dengan,

W_f = Bobot dari *forget gate*

h_{t-1} = *State* pada waktu t – 1

x_t = *Input* pada waktu t

σ = Fungsi aktivasi sigmoid

c. *Output gate*

Output gate (o_t) mengontrol seberapa banyak *state* yang berjalan ke *output* dan bekerja dengan cara yang sama dengan gerbang lainnya. Dan terakhir menghasilkan *cell state* yang baru (h_t). Rumus dari o_t dan h_t adalah

$$o_t = \sigma(W_o \times [h_{t-1}, x_t] + b_o) \quad (2.5)$$

$$h_t = o_t \times \tanh C_t \quad (2.6)$$

dengan,

W_o = Bobot dari *output gate*

h_{t-1} = *State* pada waktu t – 1

x_t = *Input* pada waktu t

σ = Fungsi aktivasi sigmoid

2.9 Evaluasi Kinerja Model

Untuk mengukur kinerja dari model yang dihasilkan salah satunya menggunakan *confusion matrix* atau juga dikenal dengan *error matrix* yaitu tabel objek spesifik yang memungkinkan visualisasi dari sebuah algoritma (Bekkar, dkk., 2013). *Confusion matrix* memiliki 4 variabel yaitu *True Positive* (TP) berarti ketika sistem memprediksi positif dan hasilnya benar positif, *True Negative* (TN) yang berarti ketika sistem memprediksi negatif dan hasilnya benar negatif, *False Positive* (FP) berarti ketika sistem memprediksi positif hasilnya salah, begitu pun dengan *False Negative* (FN) yang berarti ketika sistem memprediksi negatif dan hasilnya salah (Narkhede, 2018). Beberapa perhitungan yang biasa digunakan pada pengujian yaitu (Blessy dan Wise, 2018):

a. Akurasi adalah kedekatan antar nilai prediksi dengan nilai sebenarnya.

$$Akurasi = \frac{TP+TN}{TP+FP+TN+FN} \quad (2.7)$$

b. Presisi adalah proporsi kasus yang diprediksi positif yang juga positif benar pada data sebenarnya.

$$Presisi = \frac{TP}{TP+FP} \quad (2.8)$$

c. *Recall* adalah proporsi kasus positif yang sebenarnya diprediksi positif secara benar.

$$Recall = \frac{TP}{TP+FN} \quad (2.9)$$

d. *F1-Score* adalah perbandingan rata-rata dari presisi dan *recall*.

$$F1\ score = 2 \times \frac{recall \times presisi}{recall + presisi} \quad (2.10)$$

2.10 *K-Fold Cross Validation*

K-fold cross validation memiliki prinsip dasar membagi keseluruhan data menjadi data *training* dan data *testing*, yang berarti setiap data *training* memiliki kesempatan untuk menjadi data *testing* begitupun sebaliknya. Metode ini bekerja dengan cara mempartisi data secara acak menjadi *k* yang berukuran sama sehingga setiap bagian data dapat diprediksi jauh lebih cepat (Davidson dan Hinkley, 1997). Penentuan nilai *k* tidak terdapat aturan tertentu akan tetapi nilai *k* yang biasa digunakan yaitu 5 hingga 10 karena nilai-nilai ini ditunjukkan untuk menghasilkan estimasi tingkat kesalahan uji yang tidak mengalami bias yang terlalu tinggi maupun varians yang sangat tinggi (Govindarajan dan Chandrasekaran, 2007). Data uji pada setiap *fold* akan menghasilkan nilai akurasi model dan berlanjut ke *fold* selanjutnya sampai selesai. Jumlah total akurasi akan dibagi dengan banyaknya *k* (Yadav dan Shukla, 2016).

III. METODE PENELITIAN

3.1 Waktu dan Tempat Penelitian

3.1.1 Waktu Penelitian

Waktu penelitian pada Tabel 1 menggambarkan kegiatan penelitian dilakukan dimulai pada minggu pertama bulan November hingga minggu keempat bulan Desember 2021 dengan pengumpulan bahan materi pendukung skripsi seperti buku, jurnal, dll serta analisis data yang akan diteliti. Kemudian menyusun proposal penelitian pada bulan Januari hingga minggu ketiga Maret 2022 dan pengajuan tema proposal penelitian di minggu terakhir bulan Maret 2022. Proposal penelitian diajukan di bulan April setelah mendapat persetujuan mengenai tema penelitian. Pemaparan proposal penelitian akan dilakukan di minggu ketiga dan keempat pada bulan April 2022. Penyusunan laporan hasil penelitian dilakukan dari minggu pertama bulan Mei hingga awal minggu kedua pada bulan Agustus 2022. Pemaparan laporan hasil penelitian dilakukan pada minggu ketiga bulan Agustus 2022. Pelaksanaan ujian skripsi dilakukan pada minggu pertama bulan Oktober 2022. Waktu penelitian dapat dilihat pada Tabel 1.

Tabel 1. Tabel Waktu Penelitian

No	Jenis Kegiatan	Bulan Pelaksanaan																
		Nov		Des		Jan		Mar		Apr				Mei		Agt		Okt
		1	4	1	4	1	4	1	4	1	2	3	4	1	4	2	3	1
1	Pengumpulan bahan materi dan analisis data																	
2	Penyusunan proposal penelitian dan pengajuan judul																	
3	Pengajuan proposal																	
4	Pemaparan proposal penelitian																	
5	Penyusunan laporan hasil penelitian																	
6	Pemaparan laporan hasil penelitian																	
7	Ujian Skripsi																	

3.1.2 Tempat Penelitian

Penelitian ini dilaksanakan pada semester genap tahun ajaran 2021/2022 yang bertempat di Pusat Riset Informatika, Badan Riset dan Inovasi Nasional, Bandung.

3.2 Spesifikasi Perangkat

Penelitian ini, menggunakan dua perangkat yaitu:

a. Laptop HP-63QARV3E dengan spesifikasi sebagai berikut:

- Processor AMD Ryzen 5 5625U with Radeon Graphics 2.30 GHz
- Memori 8,00 GB RAM

- Harddisk 500 GB

b. Laptop Acer One Z1401 spesifikasi sebagai berikut:

- Processor Intel® Pentium® processor 1.7 GHz
- Memori 2,00 GB DDR L RAM
- Harddisk 200 GB

3.3 Data Penelitian

Data penelitian yang digunakan adalah data sekunder berupa judul berita *online* mengenai *covid-19* dari bulan Januari- Mei 2020 berbagai portal berita Indonesia dengan kata kunci "covid". Jenis peristiwa berita diklasifikasikan menjadi dua jenis, yaitu berita *event* diberikan kode "1" dan berita *non-event* "0". Berita *event* berarti artikel yang melaporkan peristiwa penyebaran infeksi *covid-19*. Sedangkan jenis berita *non-event* adalah artikel yang melaporkan informasi seputar pandemi *covid-19* seperti tips kesehatan, penggalangan dana untuk kelompok terdampak, dll.. Data yang terkumpul sebanyak 16.836 data dari tujuh portal berita Indonesia yang berbeda seperti "Antara", "Detik", "Kompas", "Kumparan", "Merdeka", "Republika", dan "Tempo". Data kelas minoritas (berita *event*) yang telah dipisahkan sebanyak 4.547 data. Sisanya merupakan kelas mayoritas (berita *non- event*) yaitu sebanyak 12.289 data.

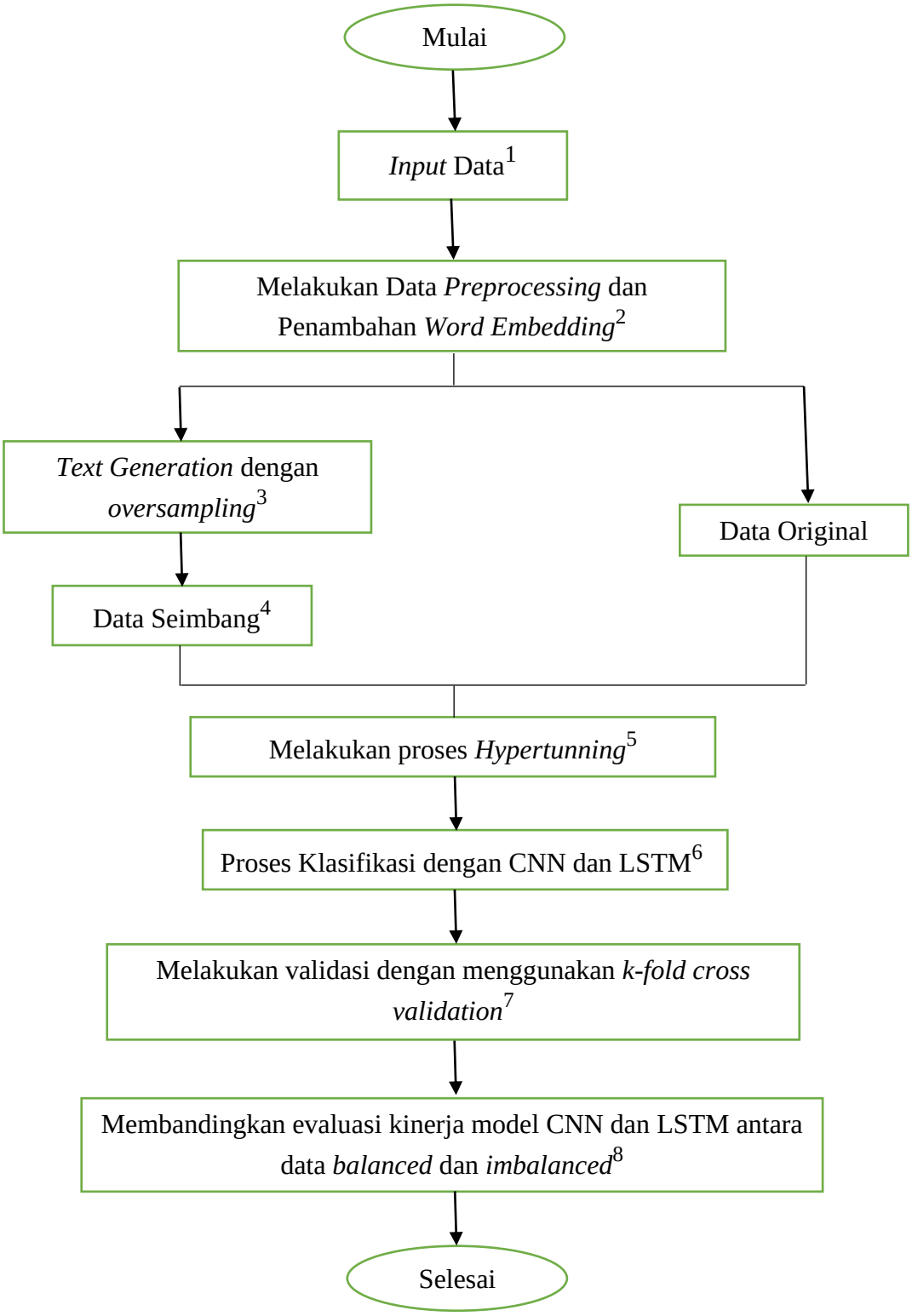
3.4 Metode Penelitian

Penelitian ini dilakukan secara studi pustaka yaitu mempelajari buku, jurnal, serta akses internet. Adapun langkah-langkah penelitian yang dilakukan sebagai berikut:

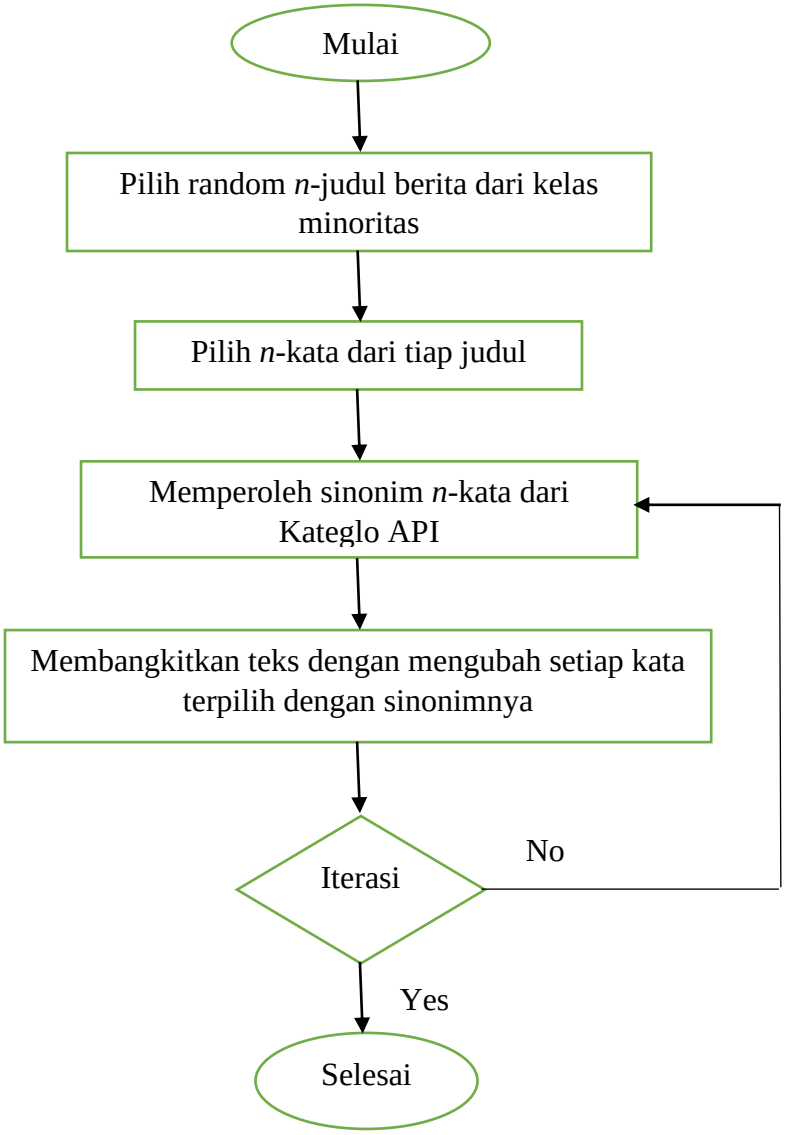
1. Mengumpulkan data judul berita *online covid-19* yang telah dilabeli sebelumnya.

2. Melakukan *pre-processing* data seperti *case folding, text cleaning, stopword*, dan *tokenizer.* Kemudian menambahkan *word embeddin*g.
3. Melakukan *n-text generation* dengan *oversampling* pada data kelas minoritas dan mencari sinonim menggunakan kamus bantuan Kateglo API.
4. Setelah dilakukan *oversampling,* data yang telah dibangkitkan akan disatukan dengan data (*non-event*) dan akan dilakukan uji klasifikasi dengan menggunakan model *deep learning*.
5. Sebelum melakukan klasifikasi, dilakukan inisialisasi parameter (*hypertunning*) pada kedua model *deep learning*. Parameter optimal yang dihasilkan pada proses *hypertunning* akan digunakan saat mengklasifikasikan kedua model *deep learning.*
6. Melakukan uji klasifikasi dengan menggunakan model *deep learning* CNN dan LSTM.
7. Melakukan validasi data dengan menggunakan *k-fold cross validation* dengan nilai *k* yaitu 1 sampai 10.
8. Setelah dilakukan uji klasifikasi, performa model data *imbalance* dan *balanced* akan dibandingkan. Metrik kinerja yang akan dibandingkan menggunakan perhitungan yang biasa digunakan pada pengujian yaitu akurasi, presisi, *recall*, dan *f1-score* (Blessy dan Wise, 2018).

   Workflow klasifikasi teks dan *n-generating text* dapat dilihat pada Gambar 4 dan 5.



Gambar 4. *Workflow* Klasifikasi Teks



Gambar 5. *Workflow n-Generating Text*

V. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil dan pembahasan mengenai model *Convolution Neural Network* (CNN) dan *Long-Short Term Memory* (LSTM) pada data *imbalanced* dan *balanced* dapat disimpulkan bahwa:

1. Hasil *hypertunning* model CNN dan LSTM, parameter optimal *epochs, batch size* dan unit LSTM pada data *balanced* memiliki nilai parameter optimal yang lebih besar daripada data *imbalanced*. Hal ini dikarenakan data *balanced* memiliki jumlah data yang lebih banyak sehingga memerlukan nilai parameter unit LSTM, *epochs* dan *batch size* yang lebih besar agar mempercepat proses untuk memperoleh hasil yang optimal.
2. Berdasarkan hasil evaluasi model CNN dan LSTM pada Gambar 10 dan 11 model CNN memiliki nilai akurasi, presisi, *recall* dan *f1-score* yang lebih baik dari model LSTM. Hal ini menunjukkan bahwa model CNN memiliki nilai ukuran metrik yang baik jika menggunakan data *imbalanced* dan *balanced*.

5.2 Saran

Adapun saran untuk penelitian selanjutnya dengan menggunakan metode *text generation* untuk mengatasi ketidakseimbangan kelas dengan menggunakan model *deep learning* yaitu:

1. Jika ingin menggunakan model LSTM untuk klasifikasi, disarankan menggunakan perangkat dengan spesifikasi RAM minimal 4 GB dan *processor* Intel® Core™ i5. Karena untuk melakukan proses *hypertunning* membutuhkan waktu komputasi yang cukup lama.

2. Hasil *text generate* dengan menggunakan sinonim kata masih memiliki *noise* dalam arti kata dan makna semantik sehingga kalimat yang diciptakan mengalami perubahan makna seperti contoh kata “misterius” berubah menjadi “sulit” seperti yang dapat dilihat pada Tabel 4. Salah satu metode yang mubngkin bisa digunakan untuk penelitian kedepan adalah metode *Generated Adversial Network* (GAN).

DAFTAR PUSTAKA

Adhi, M. S., Naf'an, M. Z., dan Usada, E. 2019. Pengaruh Sematic Expansion pada Naïve Bayes Classifier untuk Analisis Sentimen Tokoh Masyarakat. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*. **3**(2): 141-147.

Akbar, M. F., Kurniawan, I., dan Fauzi, I. 2019. Mengatasi Imbalanced Class Pada Software Defect Prediction Menggunakan Two-Step Clustering-Based Undersampling and Bagging Technique. *Jurnal Informatika*. **6**(1): 107-113.

Albewi, S., dan Mahmood, A. 2017. A Framework for Designing the Architectures of Deep Convolutional Neural Network. *Entropy*. **19**: 242

Ardiyansyah., Rahayuningsih, P. A. 2020. Penerapan Teknik Sampling Untuk Mengatasi Imbalance Class Pada Klasifikasi Online Shoppers Intention. *Jurnal Teknik Informatika Kaputama (JTIK)*. **4**(1): 7-15.

Bekkar, M. Djema, H. K., dan Alitouche, T. A. 2013. Evaluation Measures for Models Assessment over Imbalanced Data Sets. *J. Inf. Eng. Appl*. **3**(10): 27–38.

Blessy, A., dan Wise, D. D. C. J. W. 2018. Detection of Affected Part of Plant Leaves and Classification of Diseases Using CNN Technique. *International Journal of Engineering and Techniques*. **4**(2): 823-829.

Brownlee, J. 2016. How to Grid Search Parameters for Deep Learning Models in Python With Keras, 2 July 2022. https://machinelearningmastery.com/grid-search-hyperparameters-deep-learning-models-python-keras/. Diakses pada 20 Agustus 2022.

Chawla, N. V., Bowyer, K. W., Hall, L. O., dan Kegelmeyer, W. P. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*. **16**: 321–357.

Davidson, A., dan Hinkley, D. 1997. *Bootstrap Methods and their Application*. Cambridge University Press.

Gelar, T., dan Sari, A. N. 2020. Klasifikasi Komentar Video Instruksional Populer Bertemakan Pekarangan Perkotaan menggunakan Auto-Keras. *Jurnal SEICT*. **1**(1): 1-9.

Govindarajan, M., dan Chandrasekaran, M. 2007. Classifier Based Text Mining for Neural Network. *World Academy of Science, Engineering and Technology*. **27**: 200-203.

Guritno, H., dan Santosa, S. 2017. Model Klasterisasi Genre Cerpen Kompas Menggunakan K-Means. *Jurnal Teknologi Informasi*. **13**(1): 40-41.

Hao, X. Zhang, G., dan Ma, S. 2016. Deep Learning. *International Journal of Semantic Computing*. **10**(3): 417-439.

Harjanta, A. T. J. 2015. Preprocessing Text untuk Meminimalisir Kata yang Tidak Berarti dalam Proses Text Mining. *Jurnal Informatika UPGRIS*. **1**: 1-9.

Hermanto, D. T., Setyanto, A., dan Luthfi, E. T. 2021. Algoritma LSTM-CNN untuk Sentimen Klasifikasi dengan Word2vec pada Media Online. *Citec Journal*. **8**(1): 64-77.

Hijazi, S., Kumar, R., dan Rowen, C. 2015. Image Recognition Using Convolutional Neural Networks. *Cadence Whitepaper*. Hal: 1–12

Hochreiter, S., dan Schmidhuber, J. 1997. Long short-term memory. *Neural computation*. **9**(8): 1735–1780.

Kadir, A. dan Susanto, A. 2013. *Teori dan Aplikasi Pengolahan Citra*. Yogyakarta: Andi Offset.

Krizhevsky, A., Sutskever, I., dan Hinton, G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the Twenty-Sixth Annual Conference on Neural Information Processing Systems.

Mathisen, G. 2018. Forecasting Multivariate Time Series Data Using Neural Networks no. June.

Minarno, A. E., Mandiri, M. H. C., dan Alfarizy, M. R. 2021. Klasifikasi COVID-19 Menggunakan Filter Gabor dan CNN Dengan Hyperparameter Tuning. *ELKOMIKA: Jurnal Teknik Energi Elektrik, Teknik Komunikasi, dan Teknik Elektronika*. **9**(3): 403-504.

Mutmainah, S. 2021. Penanganan imbalance data pada klasifikasi kemungkinan penyakit stroke. *Jurnal Sains, Nalar, dan Aplikasi Teknologi Informasi*. **1**(1).

Naili, M., Chaibi, A. H., dan Ben Ghezala, H. H. 2017. Comparative study of word embedding methods in topic segmentation. *Procedia Computer Science*. **112**: 340–349.

Narkhede, S. 2018. Understanding Confusion Matrix, Medium, 9 May 2018. https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62/. Diakses pada 23 Maret 2022.

Nugroho, P. A., Fenriana, I., dan Arijanto, R. 2020. Implementasi Deep Learning Menggunakan Convolutional Neural Network (CNN) Pada Ekspresi Manusia. *Jurnal Algor*. **2**(1): 12-20.

Oyong, I., Utami, E., dan Luthfi, E. T. 2018. Natural language processing and lexical approach for depression symptoms screening of Indonesian twitter user, hlm. 359-364. In Proceeding 2018 10th International Conference on Information Technologies and Electrical Engineering (ICITEE).

Putra, J. W. G. 2018. *Pengenalan Pembelajaran Mesin dan Deep Learning Jan Wira Gotama Putra Pengenalan Konsep Pembelajaran Mesin dan Deep Learning*. Tokyo: Tokyo Institute of Technology.

Rohman, A. N., Utami, E., dan Raharjo, S. 2019. Deteksi Emosi Media Sosial Menggunakan Pendekatan Leksikon dan *Natural Language Processing*. *Jurnal Eksplora Informatika*. **9**(1): 70-76.

Santosa, B., dan Umam, A. 2018. *Data Mining dan Big Data Analytics*. Yogyakarta: Penebar Media Pustaka.

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.

Suartika, I. W. E. P., Wijaya, A. Y., dan Soelaiman, R. 2016. Klasifikasi Citra Menggunakan Convolutional Neural Network (CNN) Pada Caltech 101. *Jurnal Teknik ITS*. **5**(1): 65-67

Sulistiyono, M., Pristyanto, Y., dan Gumelar, G. 2021. Implementasi Algoritma Synthetic Minority Over-Sampling Technique untuk Menangani Ketidakseimbangan Kelas pada Dataset Klasifikasi. *SISTEMASI: Jurnal Sistem Informasi*. **10**(2): 445-459.

Suyanto. 2018. *Machine Learning Tingkat Dasar dan Lanjut.* Bandung: Penerbit Informatika.

Vinayakumar, R., Soman, K. P., dan Poornachandran, P. 2017. Long short-term memory based operation log anomaly detection. *International Conference on Advance Computing Communication and Informatics* (ICACCI).

Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Mahendra, R., Fung, P., dan Bahar, S. 2020. Indonlu:

Benchmark and resources for evaluating indonesian natural language understanding. *arXiv preprint arXiv:2009.05387*.

Wiranda, L., dan Sadikin, M. 2019. Penerapan Long Short Term Memory Pada Data Time Series Untuk Memprediksi Penjualan Produk PT. Metiska Farma. *Jurnal Nasional Pendidikan Teknik Informatika*. **8**(3): 184-196.

Yadav, S., dan Shukla, S. 2016. Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification. In Proceedings - 6th International Advanced Computing Conference (IACC).

Zahara, S., dan Sugianto. 2021. Peramalan Data Indeks Harga Konsumen Berbasis Time Series Multivariate Menggunakan Deep Learning. *JURNAL RESTI (Rekayasa Sistem dan Teknologi Informasi)*. **5**(1): 24-30.