IPB University
Bogor Indonesia
@Hak cipta milik IPB University
Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.
IPB University
Perpustakaan IPB University

ANALISIS SENTIMEN DARI DATA TEKS BENCANA BANJIR YANG TIDAK SEIMBANG DAN MENGANDUNG SARKASME MENGGUNAKAN METODE BERT-LSTM

NUR KHAMIDAH

PROGRAM STUDI STATISTIKA DAN SAINS DATA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
INSTITUT PERTANIAN BOGOR
BOGOR
2024

@Hak cipta milik IPB University

IPB University

IPB University
Bogor Indonesia

Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
   a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
   b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

Perpustakaan IPB University

IPB University
Bogor Indonesia

@Hak cipta milik IPB University

PERNYATAAN MENGENAI TESIS DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa tesis dengan judul "Analisis Sentimen dari Data Teks Bencana Banjir yang Tidak Seimbang dan Mengandung Sarkasme menggunakan Metode BERT-LSTM" adalah karya saya dengan arahan dari dosen pembimbing dan belum diajukan dalam bentuk apa pun kepada perguruan tinggi mana pun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka di bagian akhir tesis ini.

Dengan ini saya melimpahkan hak cipta dari karya tulis saya kepada Institut Pertanian Bogor.

Bogor, Desember 2024

Nur Khamidah
G1501221023

IPB University

Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

Perpustakaan IPB University

IPB University
Bogor Indonesia
@Hak cipta milik IPB University
Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.
IPB University
Perpustakaan IPB University

RINGKASAN

NUR KHAMIDAH. Analisis Sentimen dari Data Teks Bencana Banjir yang Tidak Seimbang dan Mengandung Sarkasme menggunakan Metode BERT-LSTM. Dibimbing oleh KHAIRIL ANWAR NOTODIPUTRO dan SACHNAZ DESTA OKTARINA.

Pengolahan bahasa alami (*Natural Language Processing*/NLP) seiring berkembangnya teknologi menjadi semakin penting, khususnya terkait dengan analisis data teks. Analisis sentimen sebagai bagian dari NLP dilakukan dengan tujuan mengidentifikasi dan mengekstraksi informasi subjektif dari teks yang diungkapkan oleh manusia, sering kali melalui metode klasifikasi yang didukung oleh data berlabel. Namun, dalam melakukan analisis sentimen dengan teknik klasifikasi kerap ditemui beberapa tantangan, antara lain terkait bagaimana klasifikasi membutuhkan label yang diperoleh melalui proses pelabelan manual yang membutuhkan biaya dan sumber daya yang besar, analisi sentimen yang berkaitan dengan topik tertentu akan cenderung mengakibatkan salah satu kelas sentimen lebih dominan sehingga mengakibatkan ketidakseimbangan kelas dalam data, serta kesulitan dalam mengidentifikasi sarkasme di mana sarkasme merupakan bagian dari tantangan dalam memahami bahasa manusia.

Penelitian sebelumnya sering menggunakan data dari platform seperti *X* (dahulu dikenal sebagai *Twitter*) untuk analisis sentimen, namun, proses pelabelan manual yang digunakan dapat menjadi pekerjaan yang rumit dan memakan waktu. Selain itu, ketidakseimbangan kelas dalam data sering kali menyebabkan masalah, di mana kelas minoritas cenderung kurang terwakili. Untuk mengatasi tantangan ini, beberapa pendekatan telah diajukan, termasuk penggunaan metode seperti SMOTE (*Synthetic Minority Over-sampling Technique*) untuk menangani ketidakseimbangan kelas dalam data. Selain itu, dalam menghadapi data teks yang memiliki struktur sekuensial, metode LSTM (*Long Short-Term Memory*) merupakan metode yang populer digunakan dalam menangani data sekuensial seperti data teks. LSTM, sebagai pengembangan dari RNN (*Recurrent Neural Networks*), mampu mengatasi masalah memori jangka panjang dan lebih efektif dalam mengingat informasi dalam jangka waktu yang lebih lama. Tantangan lain yang sering dihadapi dalam analisis sentimen adalah mengenali sarkasme, sebuah bentuk ekspresi yang sulit untuk diinterpretasikan bahkan oleh manusia. Model bahasa BERT (*Bidirectional Encoder Representations from Transformers*) telah terbukti efektif dalam menganalisis sentimen. Dengan metode *deep learning* secara *bidirectional*, BERT dapat memahami konteks bahasa alami dengan lebih baik.

Penelitian ini bertujuan untuk menguji kinerja beberapa model klasifikasi, yaitu BERT, LSTM, dan model kombinasi BERT-LSTM, dalam mengenali teks yang mengandung sarkasme, khususnya dalam konteks banjir di Indonesia. Melalui penggunaan teknik penanganan data tidak seimbang SMOTE, *Random Undersampling*, dan SMOTETomek, penelitian ini juga mengevaluasi apakah penanganan data tidak seimbang tersebut mampu meningkatkan kinerja model klasifikasi. Pemodelan menggunakan BERT, LSTM, dan BERT-LSTM dilakukan dengan terlebih dahulu melakukan *tuning hyperparameter* antara lain *epochs, batch size*, dan *learning rate*. Mengingat komputasi yang dilakukan berat dan tidak memungkinkan untuk menggunakan prosesor CPU, digunakan bantuan prosesor

IPB University
Bogor Indonesia

lain seperti GPU T4, NVIDIA GPU A100 dan V100 untuk menjalankan model. Seluruh prosedur pemodelan ini dilakukan melalui *Google Colaboratory* dengan bahasa pemrograman Python 3.10.6.

Terdapat dua jenis data yang digunakan dalam penelitian ini, antara lain data simulasi teks dan data empiris. Data simulasi digenerasi melalui ChatGPT 3.5 dengan *prompt* atau instruksi tertentu dan dikelompokkan menjadi tiga jenis label antara lain SARKAS (data teks mengandung sarkasme dengan jelas), BETWEEN (data teks kurang jelas terdefiisi sebagai sarkasme atau bukan sarkasme, dan NONSARKAS (data teks dengan jelas terdefinisi tidak mengandung sarkasme) dengan muatan *corpus* atau *term* yang disesuaikan dengan data empiris yang sebelumnya telah dieksplorasi sehingga struktur kata yang dihasilkan lebih sesuai dengan data aslinya. Pemodelan pada data simulasi dimaksudkan untuk mengidentifikasi apakah model yang digunakan mampu mengenali data teks yang mengandung sarkasme dengan baik. Data empiris diperoleh dengan metode *scraping* melalui media sosial *Twitter/X* dengan X API v.2 pada rentang waktu 19 Desember 2023 hingga 13 Januari 2024 dan diperoleh total 10.479 *post.* Data empiris yang diperoleh selanjutnya dilakukan pembersihan dan pelabelan oleh 9 (sembilan) orang *annotator* yang berasal dari 3 (tiga) bidang antara lain Statistika, Bahasa dan Sastra Indonesia, serta Psikologi untuk mengetahui apakah terdapat pengaruh yang diberikan oleh perbedaan bidang *annotator* terhadap hasil pelabelan dan performa prediksi model. Seluruh data ini selanjutnya dilakukan eksplorasi, pra-pemrosesan, dan vektorisasi menggunakan TF-IDF sebagaimana pada pemrosesan data teks pada umumnya.

Hasil dari penelitian ini menunjukkan bahwa model BERT dan BERT-LSTM mampu mengklasifikasikan teks mengandung sarkasme dengan baik dan akurat, baik dengan atau tanpa dilakukan teknik penanganan ketidakseimbangan kelas dalam data dengan nilai akurasi terboboti masing-masing secara berturut-turut sebesar 99,96% dan 99,97% pada data latih dan sebesar 98,7% dan 97,8% pada data uji. Selain itu, pemodelan pada data empiris menunjukkan bahwa model BERT dan BERT-LSTM memiliki kinerja yang lebih baik dibandingkan dengan model LSTM dalam mengklasifikasikan sentimen teks terkait banjir dengan nilai akurasi terboboti yang diperoleh lebih besar dan stabil pada semua jenis data berdasarkan bidang *annotator* yang dimodelkan. Penelitian ini memberikan wawasan tambahan tentang kemampuan model-model tersebut dalam mengatasi tantangan dalam analisis sentimen, serta pemahaman lebih lanjut tentang persepsi masyarakat terhadap bencana banjir di Indonesia.

Kata kunci: sentimen, banjir, sarkasme, data tidak seimbang, BERT, LSTM

@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

IPB University

Perpustakaan IPB University

IPB University
Bogor Indonesia
@Hak cipta milik IPB University
Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.
IPB University
Perpustakaan IPB University

SUMMARY

NUR KHAMIDAH. Sentiment Analysis of Flood Disaster Text Data Suffering from Imbalance Problems and Containing Sarcasm using BERT-LSTM Method. Supervised by KHAIRIL ANWAR NOTODIPUTRO and SACHNAZ DESTA OKTARINA.

Natural Language Processing (NLP) is becoming increasingly important as technology evolves, especially in relation to text data analysis. Sentiment analysis, as part of NLP, is conducted with the aim of identifying and extracting subjective information from text expressed by humans, often through classification methods supported by labelled data. However, performing sentiment analysis with classification techniques often encounters several challenges, including the need for labels obtained through manual labelling, which requires significant costs and resources. Sentiment analysis related to specific topics tends to result in one sentiment class being more dominant, leading to class imbalance in the data, and difficulties in identifying sarcasm, which is part of the challenge of understanding human language.

Previous research often used data from platforms like X (formerly known as Twitter) for sentiment analysis. However, the manual labelling process used can become complicated and time-consuming. Additionally, class imbalance in the data often causes problems, with minority classes tending to be underrepresented. To address these challenges, several approaches have been proposed, including the use of methods like SMOTE (Synthetic Minority Over-sampling Technique) to handle class imbalance in data. Furthermore, in dealing with text data that has a sequential structure, the LSTM (Long Short-Term Memory) method is popularly used for handling sequential data like text. LSTM, as a development of RNN (Recurrent Neural Networks), can overcome long-term memory issues and is more effective in retaining information over a longer period. Another challenge often faced in sentiment analysis is recognizing sarcasm, a form of expression that is difficult to interpret even by humans. The BERT (Bidirectional Encoder Representations from Transformers) language model has proven effective in sentiment analysis. With deep learning methods in a bidirectional manner, BERT can understand natural language context better.

This study aims to test the performance of several classification models, namely BERT, LSTM, and the combined BERT-LSTM model, in recognizing text containing sarcasm, particularly in the context of flooding in Indonesia. Using data imbalance handling techniques such as SMOTE, Random Undersampling, and SMOTETomek, this study also evaluates whether these data imbalance handling techniques can improve the performance of classification models. Modelling using BERT, LSTM, and BERT-LSTM is carried out by first tuning hyperparameters, including epochs, batch size, and learning rate. Considering the heavy computation involved, which is not feasible using a CPU processor, other processors like GPU T4, NVIDIA GPU A100, and V100 are used to run the models. All modelling procedures are carried out through Google Colaboratory with Python 3.10.6 programming language.

There are two types of data used in this study, namely simulation text data and empirical data. Simulation data is generated through ChatGPT 3.5 with specific

prompts or instructions and grouped into three types of labels: SARKAS (text data clearly containing sarcasm), BETWEEN (text data less clearly defined as sarcasm or non-sarcasm), and NONSARKAS (text data clearly defined as non-sarcasm), with corpus or term content adjusted to previously explored empirical data to make the word structure produced more in line with the original data. Modeling on simulation data is intended to identify whether the models used can recognize text data containing sarcasm well. Empirical data is obtained through scraping from social media Twitter/X using X API v.2 between December 19, 2023, and January 13, 2024, with a total of 10,479 posts. The empirical data obtained is then cleaned and labelled by 9 annotators from three fields: Statistics, Indonesian Language and Literature, and Psychology to determine whether the annotator's field influences the labelling results and model prediction performance. All this data is then explored, pre-processed, and vectorized using TF-IDF, as in general text data processing.

The results of this study show that the BERT and BERT-LSTM models can classify text containing sarcasm accurately, both with or without class imbalance handling techniques, with weighted accuracy values 99.96% dan 99.97% on train data and 98.7% dan 97.8% on test data, respectively. Additionally, modelling on empirical data shows that the BERT and BERT-LSTM models perform better than the LSTM model in classifying text sentiment related to flooding, with higher and more stable weighted accuracy values across all data types based on the annotator's field. This study provides additional insights into the capability of these models in addressing challenges in sentiment analysis and further understanding of public perception towards flooding disasters in Indonesia.

Keywords: *sentiment, flood, sarcasm, imbalance data, BERT, LSTM*

IPB University
Bogor Indonesia
@Hak cipta milik IPB University
IPB University
Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.
Perpustakaan IPB University

IPB University
Bogor Indonesia
@Hak cipta milik IPB University
IPB University
Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.
Perpustakaan IPB University

© Hak Cipta milik IPB, tahun 2024
Hak Cipta dilindungi Undang-Undang

Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan atau menyebutkan sumbernya. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik, atau tinjauan suatu masalah, dan pengutipan tersebut tidak merugikan kepentingan IPB.

Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apa pun tanpa izin IPB.

IPB University
Bogor Indonesia
@Hak cipta milik IPB University
Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.
IPB University
Perpustakaan IPB University

ANALISIS SENTIMEN DARI DATA TEKS BENCANA BANJIR YANG TIDAK SEIMBANG DAN MENGANDUNG SARKASME MENGGUNAKAN METODE BERT-LSTM

NUR KHAMIDAH

Tesis
sebagai salah satu syarat untuk memperoleh gelar
Magister Sains pada
Program Studi Statistika dan Sains Data

PROGRAM STUDI STATISTIKA DAN SAINS DATA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
INSTITUT PERTANIAN BOGOR
BOGOR
2024

IPB University
Bogor Indonesia
@Hak cipta milik IPB University
IPB University
Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.
Perpustakaan IPB University

Tim Penguji pada Ujian Tesis:

1. Dr. Yenni Angraini, S.Si., M.Si.

IPB University
Bogor Indonesia
@Hak cipta milik IPB University
Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.
Perpustakaan IPB University

Judul Tesis : Analisis Sentimen dari Data Teks Bencana Banjir yang Tidak Seimbang dan Mengandung Sarkasme menggunakan Metode BERT-LSTM
Nama : Nur Khamidah
NIM : G1501221023

Disetujui oleh

Pembimbing 1:
Prof. Dr. Ir. Khairil Anwar Notodiputro, M.S.

Pembimbing 2:
Sachnaz Desta Oktarina, S.Stat., M.Agr., Ph.D.

Diketahui oleh

Ketua Program Studi:
Dr. Agus Mohamad Soleh
NIP 197503151999031000

Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam:
Dr. Berry Juliandi, S.Si., M.Si.
NIP 197807232007011001

Tanggal Ujian:
22 Oktober 2024

Tanggal Lulus:

IPB University
Bogor Indonesia
Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.
@Hak cipta milik IPB University
IPB University
Perpustakaan IPB University

PRAKATA

Puji dan syukur penulis panjatkan kepada Allah subhaanaahu wa ta'ala atas limpahan nikmat dan karunia-Nya yang tiada henti-hentinya sehingga tesis berjudul "Analisis Sentimen dari Data Teks Bencana Banjir yang Tidak Seimbang dan Mengandung Sarkasme menggunakan Metode BERT-LSTM" ini berhasil diselesaikan. Shalawat serta salam semoga senantiasa tercurahkan kepada baginda Nabi Muhammad ﷺ serta keluarga, sahabat, dan pengikutnya hingga akhir zaman. Tesis ini merupakan salah satu syarat untuk memperoleh gelar Magister Sains pada Program Studi Statistika Departemen Statistika Institut Pertanian Bogor.

Selama proses penyusunan tesis ini, berbagai bantuan dan dukungan telah penulis dapatkan sehingga proses penyusunan tesis ini telah penulis selesaikan. Penulis mengucapkan terima kasih kepada pihak-pihak berikut:

1. Bapak Prof. Dr. Ir. Khairil Anwar Notodiputro, M.S. dan Ibu Sachnaz Desta Oktarina, S.Stat., M.Agr., Ph.D. selaku komisi pembimbing tesis yang senantiasa membantu penulis dengan memberikan masukan, saran, dan arahan yang sangat berarti dalam penyusunan tesis ini.
2. Seluruh pengajar sekolah pascasarjana Departemen Statistika IPB yang telah banyak memberikan ilmu selama perkuliahan sampai dengan penyusunan tesis ini.
3. Keluarga penulis, Bapak Ahmad Khoiri, ibu Siti Ulfah, Kakak Fathul Amin, dan Kakak Khoirotul Islakhiyah yang senantiasa memberikan dukungan dan doa pada setiap proses yang dilampaui penulis selama penyelesaian tesis.
4. Seluruh Staf Tata Usaha Departemen Statistika IPB yang membantu segala proses administrasi.
5. Teman-teman mahasiswa Himpunan Profesi Pascasarjana Statistika IPB, teman-teman mahasiswa Angkatan 2022 Pascasarjana Statistika IPB, terkhusus kepada Dhea, Akmarina, Fatma, Tiya, Mbak Vivin, dan Mas Teguh atas dukungan dan semangat yang diberikan kepada penulis dalam menyelesaikan tesis ini.
6. Lembaga Pengelola Dana Pendidikan Kementerian Keuangan Republik Indonesia atas pembiayaan dan bantuan material yang diberikan serta segenap *awardee* di Kelurahan IPB.

Semoga tesis memberikan banyak manfaat. Mohon maaf jika terdapat kesalahan dalam penyusunan tesis ini. Semoga Allah SWT selalu melimpahkan rahmat serta hidayah-Nya kepada kita semua. Aamin.

Bogor, Desember 2024

Nur Khamidah

IPB University
Bogor Indonesia
@Hak cipta milik IPB University
Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.
IPB University
Perpustakaan IPB University

DAFTAR ISI

HALAMAN SAMPUL		i
LAMAN PERNYATAAN DAN PELIMPAHAN HAK CIPTA		iii
RINGKASAN		iv
SUMMARY		vi
HALAMAN HAK CIPTA		viii
HALAMAN JUDUL DALAM		ix
HALAMAN TIM PENGUJI		x
HALAMAN PENGESAHAN		xi
PRAKATA		xii
DAFTAR ISI		xiii
DAFTAR TABEL		i
DAFTAR GAMBAR		i
DAFTAR LAMPIRAN		iii
I	PENDAHULUAN	1
1.1	Latar Belakang	1
1.2	Rumusan Masalah	3
1.3	Tujuan	3
1.4	Manfaat	3
II	TINJAUAN PUSTAKA	5
2.1	Analisis Sentimen	5
2.2	Sarkasme	5
2.3	Pemrosesan Data Teks	6
2.4	Latent Dirichlet Allocation	7
2.5	*Bidirectional Encoder Representations from Transformers* (BERT)	9
2.6	*Long Short-Term Memory* (LSTM)	10
2.7	Penanganan Data Tidak Seimbang	13
2.8	Evaluasi Kebaikan Klasifikasi	14
III	METODE	20
3.1	Data	20
3.2	Prosedur Analisis Data	22
IV	HASIL DAN PEMBAHASAN	26
4.1	Kajian Simulasi	26
4.2	Penerapan pada Data Empiris	32
V	PENUTUP	52
5.1	Kesimpulan	52
5.2	Saran	52
DAFTAR PUSTAKA		54
LAMPIRAN		59
RIWAYAT HIDUP		122

@Hak cipta milik IPB University

IPB University

IPB University
Bogor Indonesia

Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
   a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
   b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

Perpustakaan IPB University

IPB University
Bogor Indonesia
@Hak cipta milik IPB University
Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.
Perpustakaan IPB University

DAFTAR TABEL

1	Contoh label teks untuk analisis setimen	5
2	Matriks konfusi untuk peubah respon lebih dari dua kelas (Markoulidakis *et al.* 2021)	15
3	Contohi/ilustrasi matriks konfusi pada klasifikasi multikelas	16
4	Data simulasi teks	21
5	Parameter dan nilai yang digunakan dalam *crawling* data	21
6	Spesifiksi prosesor yang digunakan dalam penelitian	22
7	Hasil nilai akurasi terboboti prediksi data simulasi pada data latih	27
8	Hasil nilai akurasi terboboti prediksi data simulasi pada data uji	28
9	Tabel analisis ragam nilai akurasi terboboti pada data simulasi	29
10	Hasil uji ANOVA dan uji lanjut BNJ Tukey berdasarkan perbedaan teknik penanganan data tidak seimbang di masing-masing model pada data simulasi	30
11	Hasil uji ANOVA dan uji lanjut BNJ Tukey berdasarkan perbedaan model di masing-masing teknik penanganan data tidak seimbang pada data simulasi	30
12	Contoh data teks yang diambil	33
13	Data *tweets* setelah dilakukan *case folding*	33
14	Data *tweets* setelah dilakukan *filtering*	34
15	Data *tweets* setelah dilakukan penghapusan *stopwords*	35
16	Data *tweets* setelah dilakukan *stemming*	35
17	Data *tweets* setelah dilakukan *stemming*	35
18	Hasil perankingan dengan *TF-IDF Vectorizer* pada lima term teratas	36
19	Topik yang terbentuk dan sebaran kata dari masing-masing topik	39
20	Topik yang terbentuk dan sebaran kata dari masing-masing topik	40
21	Nilai *hyperparameter* yang di-*tuning*	41
22	Hasil pencarian *hyperparameter* terbaik dengan *grid search*	42
23	Hasil klasifikasi data empiris berdasarkan nilai akurasi terboboti pada data latih	45
24	Hasil klasifikasi data empiris berdasarkan nilai akurasi terboboti pada data uji	46
25	Tabel analisis ragam nilai akurasi terboboti pada klasifikasi data simulasi	49
26	Hasil uji BNJ Tukey berdasarkan perbedaan model pada analisis ragam data empiris	49
27	Hasil uji BNJ Tukey berdasarkan perbedaan teknik penanganan data tidak seimbang pada analisis ragam data empiris	50

DAFTAR GAMBAR

1	Banjir menjadi *trend* di Twitter.	3
2	Grafik representasi LDA (Blei *et al.* 2003)	8
3	Arsitektur BERT (Devlin *et al.* 2019)	9
4	Sel RNN (Yu *et al.* 2019)	10

IPB University

5	LSTM tanpa *forget gate* (Yu *et al.* 2019)	11
6	LSTM dengan *forget gate* (Yu *et al.* 2019)	12
7	Sebaran data simulasi	20
8	Alur analisis	25
9	Perbandingan hasil awan kata (*word cloud*) dari kedua jenis data: (i) data simulasi (ii) data empiris pada *post* yang mengandung sarkasme	26
10	Hasil kinerja klasifikasi berdasarkan nilai akurasi terboboti pada data simulasi	27
11	Hasil evaluasi durasi *running* pada masing-masing prosesor pada pemodelan data simulasi	28
12	Hasil uji lanjut pada interaksi model dan teknik penanganan data tidak seimbang yang digunakan	31
13	Matriks pembobot TFIDF	36
14	Hasil awan kata data teks yang telah melalui proses *cleaning*	37
15	Hasil awan kata data teks pada label “negatif”	38
16	Hasil awan kata data teks pada label “netral”	38
17	Hasil awan kata data teks pada label “negatif”	38
18	Hasil pemodelan LDA pada *n* topik yang berbeda	39
19	Hasil pemodelan LDA pada *n* topik yang berbeda pada masing-masing kelas	40
20	Sebaran proporsi masing-masing kelas pada setiap data	41
21	Hasil pemodelan klasifikasi pada data empiris hasil *ensembling* tiga bidang *annotator*	42
22	Hasil pemodelan klasifikasi pada data empiris yang dilabeli *annotator* bidang Statistika	43
23	Hasil pemodelan klasifikasi pada data empiris yang dilabeli *annotator* bidang Bahasa/Sastra Indonesia	43
24	Hasil pemodelan klasifikasi pada data empiris yang dilabeli *annotator* bidang Psikologi	43
25	Hasil evaluasi durasi *running* pada masing-masing prosesor pada pemodelan data empiris	47
26	Hasil uji lanjut data empiris berdasarkan pengaruh pemilihan model klasifikasi (kiri) dan teknik penanganan data tidak seimbang (kanan)	51

IPB University
Bogor Indonesia
@Hak cipta milik IPB University
IPB University
Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.
Perpustakaan IPB University

IPB University
Bogor Indonesia

@Hak cipta milik IPB University

DAFTAR LAMPIRAN

1	Proses pembangkitan data simulasi teks melalui ChatGPT 3.5	60
2	Hasil tabel kontingensi pada masing-masing *annotator* dan bidangnya	63
3	*Source code* yang dijalankan selama proses penelitian	65

Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

IPB University

Perpustakaan IPB University