PENGGUNAAN MODEL BERT UNTUK KLASIFIKASI TEKS DAN DETEKSI ENTITAS DATASET BAHAYA PRODUK MAKANAN

Laporan Tugas Akhir

Disusun sebagai syarat kelulusan tingkat sarjana

Oleh

M FARREL DANENDRA RACHIM

NIM : 13521048

PROGRAM STUDI TEKNIK INFORMATIKA

SEKOLAH TEKNIK ELEKTRO DAN INFORMATIKA

INSTITUT TEKNOLOGI BANDUNG

Maret 2025

PENGGUNAAN MODEL BERT UNTUK KLASIFIKASI TEKS DAN DETEKSI ENTITAS DATASET BAHAYA PRODUK MAKANAN

Laporan Tugas Akhir

Oleh

M FARREL DANENDRA RACHIM

NIM : 13521048

Program Studi Teknik Informatika

Sekolah Teknik Elektro dan Informatika

Institut Teknologi Bandung

Telah disetujui dan disahkan sebagai Laporan Tugas Akhir

di Bandung, pada tanggal <tanggal>

Pembimbing,

<Nama dan Gelar Pembimbing>

NIP <NIP Pembimbing>

PENGGUNAAN MODEL BERT UNTUK KLASIFIKASI TEKS DAN DETEKSI ENTITAS DATASET BAHAYA PRODUK MAKANAN

Laporan Tugas Akhir

Oleh

M FARREL DANENDRA RACHIM

NIM : 13521048

Program Studi Teknik Informatika

Sekolah Teknik Elektro dan Informatika

Institut Teknologi Bandung

Telah disetujui dan disahkan sebagai Draft Laporan Tugas Akhir

di Bandung, pada tanggal <tanggal>

Pembimbing I,	Pembimbing II,
<Nama dan Gelar Pembimbing I>	<Nama dan Gelar Pembimbing II>
NIP <NIP Pembimbing I>	NIP <NIP Pembimbing II>

LEMBAR PERNYATAAN

Dengan ini saya menyatakan bahwa:

1. Pengerjaan dan penulisan Laporan Tugas Akhir ini dilakukan tanpa menggunakan bantuan yang tidak dibenarkan.
2. Segala bentuk kutipan dan acuan terhadap tulisan orang lain yang digunakan di dalam penyusunan laporan tugas akhir ini telah dituliskan dengan baik dan benar.
3. Laporan Tugas Akhir ini belum pernah diajukan pada program pendidikan di perguruan tinggi mana pun.

Jika terbukti melanggar hal-hal di atas, saya bersedia dikenakan sanksi sesuai dengan Peraturan Akademik dan Kemahasiswaan Institut Teknologi Bandung bagian Penegakan Norma Akademik dan Kemahasiswaan khususnya Pasal 2.1 dan Pasal 2.2.

Bandung, <tanggal>

M Farrel Danendra Rachim
NIM 13521048

KATA PENGANTAR

Gunakan bagian ini untuk memberikan ucapan terima kasih kepada semua pihak yang secara langsung atau tidak langsung membantu penyelesaian tugas akhir, termasuk pemberi beasiswa jika ada. Utamakan untuk memberikan ucapan terima kasih kepada tim pembimbing tugas akhir dan staf pengajar atau pihak program studi, bahkan sebelum mengucapkan terima kasih kepada keluarga. Ucapan terima kasih sebaiknya bukan hanya menyebutkan nama orang saja, tetapi juga memberikan penjelasan bagaimana bentuk bantuan/dukungan yang diberikan. Gunakan bahasa yang baik dan sopan serta memberikan kesan yang enak untuk dibaca. Sebagai contoh: “Tidak lupa saya ucapkan terima kasih kepada teman dekat saya, Tito, yang sejak satu tahun terakhir ini selalu memberikan semangat dan mengingatkan saya apabila lengah dalam mengerjakan Tugas Akhir ini. Tito juga banyak membantu mengoreksi format dan *layout* tulisan. Apresiasi saya sampaikan kepada pemberi beasiswa, Yayasan Beasiswa, yang telah memberikan bantuan dana kuliah dan biaya hidup selama dua tahun. Bantuan dana tersebut sangat membantu saya untuk dapat lebih fokus dalam menyelesaikan pendidikan saya. ....”. Ucapan permintaan maaf karena kekurangsempurnaan hasil Tugas Akhir tidak perlu ditulis.

DAFTAR ISI

DAFTAR LAMPIRAN

DAFTAR GAMBAR

DAFTAR TABEL

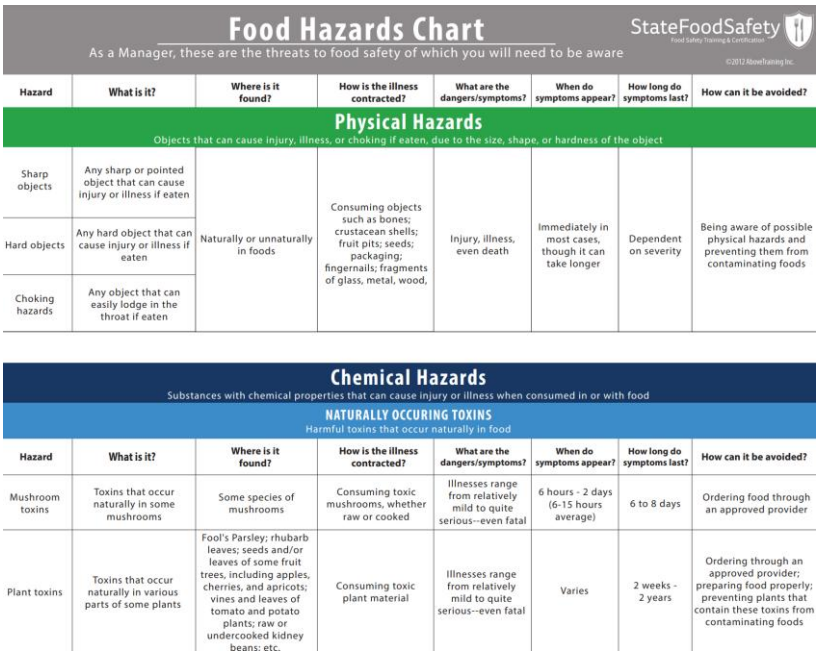
BAB I

PENDAHULUAN

I.1 Latar Belakang

Bahaya makanan atau *food hazard* adalah sebuah agen yang memberi dampak negatif terhadap kesehatan manusia (Singh dkk.., 2019). Konsumsi makanan yang tidak aman yang mengandung bahaya makanan seperti bakteri berbahaya, virus, parasit, dan zat kimia dapat berdampak signifikan pada kesehatan individu. Dengan ribuan catatan bahaya baru yang ditambahkan setiap tahun, menjadi semakin sulit untuk melacak temuan-temuan baru dan potensi risiko terkait keamanan makanan. Pemantauan bahaya makanan di seluruh rantai pasok makanan menjadi penting untuk menjamin berjalannya sistem manajemen keamanan makanan dengan baik (ISO, 2013).

Terdapat beberapa jenis bahaya makanan tergolong berdasarkan wujudnya, seperti *physical hazard, chemical hazard, biological hazard,* dan lain-lain. Berikut adalah cuplikan tabel yang berisi penjelasan mengenai jenis bahaya makanan bersifat fisik dan kimiawi (StateFoodSafety, diakses 2024).



Food Hazards Chart

As a Manager, these are the threats to food safety of which you will need to be aware

StateFoodSafety

©2012 AboveTraining Inc.

Hazard	What is it?	Where is it found?	How is the illness contracted?	What are the dangers/symptoms?	When do symptoms appear?	How long do symptoms last?	How can it be avoided?
Physical Hazards — Objects that can cause injury, illness, or choking if eaten, due to the size, shape, or hardness of the object							
Sharp objects	Any sharp or pointed object that can cause injury or illness if eaten	Naturally or unnaturally in foods	Consuming objects such as bones; crustacean shells; fruit pits; seeds; packaging; fingernails; fragments of glass, metal, wood,	Injury, illness, even death	Immediately in most cases, though it can take longer	Dependent on severity	Being aware of possible physical hazards and preventing them from contaminating foods
Hard objects	Any hard object that can cause injury or illness if eaten						
Choking hazards	Any object that can easily lodge in the throat if eaten						
Chemical Hazards — Substances with chemical properties that can cause injury or illness when consumed in or with food							
NATURALLY OCCURING TOXINS — Harmful toxins that occur naturally in food							
Mushroom toxins	Toxins that occur naturally in some mushrooms	Some species of mushrooms	Consuming toxic mushrooms, whether raw or cooked	Illnesses range from relatively mild to quite serious--even fatal	6 hours - 2 days (6-15 hours average)	6 to 8 days	Ordering food through an approved provider
Plant toxins	Toxins that occur naturally in various parts of some plants	Fool's Parsley; rhubarb leaves; seeds and/or leaves of some fruit trees, including apples, cherries, and apricots; vines and leaves of tomato and potato plants; raw or undercooked kidney beans; etc.	Consuming toxic plant material	Illnesses range from relatively mild to quite serious--even fatal	Varies	2 weeks - 2 years	Ordering through an approved provider; preparing food properly; preventing plants that contain these toxins from contaminating foods

Gambar I.1 Tabel Penjelasan Bahaya Makanan Fisik dan Kimiawi

Pembelajaran mesin (*Machine Learning*/ML) untuk tujuan pemantauan dan prediksi keamanan makanan telah digunakan dalam berbagai penelitian (Wang dkk.., 2022; Wang dkk.., 2023). Untuk metode pembelajaran mesin ini, akan digunakan model *Bidirectional Encoder Representations from Transformers* (BERT). BERT dirancang untuk melakukan *pretraining* representasi *bidirectional* yang mendalam dari teks tanpa label. Sebagai hasilnya, model BERT yang telah dilatih sebelumnya dapat disesuaikan (*fine-tuned*) hanya dengan menambahkan satu lapisan output tambahan untuk menciptakan model mutakhir untuk berbagai jenis tugas, salah satunya adalah klasifikasi dan ekstraksi informasi teks (Devlin dkk., 2019).

Namun, penerapan dan eksplorasi dalam klasifikasi bahaya makanan berbasis teks masih sangat terbatas, khususnya penelitian yang menggunakan BERT. Hal ini disebabkan oleh perkembangan metode ML yang sangat cepat, serta data keamanan

makanan yang tersebar di berbagai domain yang kurang langsung berkaitan dengan keamanan makanan (Marvin, Bouzembrak, Janssen dkk.., 2017). Kendala di atas juga berlaku dalam proses pencarian entitas bahaya makanan, yaitu *item* bahaya dan makanan pada dataset secara spesifik, bukan dalam sebuah kelas atau golongan tertentu. Selain itu, model konvensional seperti Bayesian Network (BN) kurang mampu memproses data tekstual yang menjadi data mentah dari log bahaya makanan yang sudah ada. Model-model konvensional seperti ini lebih cocok digunakan untuk memprediksi data makanan berdasarkan label target yang sudah digolongkan sebelumnya tanpa konten tekstual. Sebagai contoh, telah dilakukan studi klasifikasi yang menggunakan BN untuk mengukur korelasi antara bahaya keamanan makanan penelitian yang tidak mengandalkan data teks sama sekali (Bouzembrak & Marvin, 2019). BN yang digunakan memiliki kinerja yang baik (95%), namun BERT telah terbukti lebih cocok digunakan dalam tugas klasifikasi dan ekstraksi yang sangat dependen terhadap hal-hal seperti tokenisasi dan pengenalan nama entitas. Sebuah studi yang memanfaatkan model BERT dan BioBERT membuktikan bahwa performa model dalam melakukan ekstraksi informasi makanan mampu meraih nilai yang tinggi, yakni sekitar 73.39% sampai 78.96%. (Stojanov dkk, 2021).

Terdapat potensi bahwa pembelajaran entitas dan klasifikasi teks makanan menggunakan BERT dapat berjalan dengan efektif berdasarkan studi-studi yang telah dilakukan sebelumnya. BERT merupakan model yang mudah dilatih oleh data *pre-training* dan *fine-tuning* sehingga dapat melakukan prediksi label target lebih akurat dibandingkan model ML lainnya seperti Random Forest, BN, atau SVM. Eksplorasi lebih lanjut dalam topik ini dapat membantu manusia dengan cepat mengevaluasi validitas prediksi terkait bahaya makanan dan produk yang bersangkutan. Selain itu, BERT dapat diadaptasi dengan *fine-tuning* pada dataset spesifik, memungkinkan peningkatan akurasi untuk tugas-tugas spesifik seperti klasifikasi jenis bahaya makanan dan ekstraksi entitas. Model ini bisa memanfaatkan data yang ada dengan lebih efisien dibandingkan harus membangun model dari nol. Performa model BERT juga dapat ditingkatkan dengan melakukan

tahap *pre-processing* terhadap data yang telah tersedia, seperti augmentasi data teks, serta mengatur *hyperparameter* model BERT tersebut untuk memperoleh hasil evaluasi maksimal, seperti jumlah *epoch*.

Studi ini menyediakan sebuah alternatif untuk mengklasifikasikan dan melakukan ekstraksi entitas bahaya makanan yang melibatkan BERT. Model BERT ini akan dilatih menggunakan data bahaya dan produk makanan yang telah diolah melalui *pre-processing*. Sistem klasifikasi teks dan deteksi entitas ini mampu mendeteksi entitas dan bahaya makanan dengan tepat serta mencocokkannya dengan label yang sudah tersedia di data latih. Sistem ini memiliki dua tugas utama: (1) Klasifikasi teks untuk prediksi bahaya makanan, yaitu memprediksi jenis bahaya dan produk, serta (2) Deteksi entitas bahaya makanan dan produk, yaitu memprediksi bahaya dan produk secara spesifik.

I.2 Rumusan Masalah

Berdasarkan latar belakang yang sudah dijelaskan, berikut adalah rumusan masalah yang akan dibahas pada Tugas Akhir ini.

1. Bagaimana cara mengklasifikasikan teks bahaya makanan (*food hazard*) menggunakan BERT, agar memperoleh klasifikasi tipe bahaya dan produk makanan secara umum?
2. Bagaimana cara mendeteksi entitas bahaya makanan dan produk menggunakan BERT, agar memperoleh bahaya dan produk makanan secara spesifik?

I.3 Tujuan

Berikut adalah tujuan yang akan dicapai pada Tugas Akhir ini berdasarkan uraian latar belakang dan rumusan masalah.

1. Membangun model BERT yang dapat melakukan klasifikasi akurat untuk bahaya dan produk makanan.

2. Membangun model BERT yang dapat melakukan deteksi entitas bahaya dan produk makanan.

I.4 Batasan Masalah

Berikut adalah batasan masalah pada Tugas Akhir ini.

1. Tugas klasifikasi teks dan deteksi entitas serta dataset yang digunakan memiliki bahasa Inggris.
2. Klasifikasi teks dan deteksi entitas dilakukan menggunakan model *pre-trained* "bert-base-uncased" dengan mempertimbangkan keterbatasan sumber daya yang tersedia.
3. Model untuk NER hanya digunakan untuk mengekstraksi entitas produk saja, bukan entitas *hazard*.

Commented [MR1]: Apa ada lagi batasannya?

I.5 Metodologi

Berikut adalah metodologi yang digunakan dalam pengembangan Tugas Akhir ini.

1. Pemrosesan awal data (*pre-processing*)

   Pada tahap awal, dilakukan *pre-processing* data untuk menambahkan variasi terhadap data yang sudah ada dengan metode augmentasi teks. Hal ini dilakukan agar performa model meningkat akibat variasi kata yang diterima.

2. Implementasi dan eksperimen

   Pada tahap kedua, dilakukan implementasi solusi dan eksperimen terhadap model dan *hyperparameter* yang digunakan untuk melatih dataset. Proses ini diadakan untuk mengamati model apa dan *hyperparameter* apa saja yang mampu mencapai nilai optimal bagi model yang sudah ada.

3. Evaluasi hasil

   Pada tahap ketiga, dilakukan evaluasi terhadap eksperimen yang telah dilakukan menggunakan metrik yang sudah ditentukan sebelumnya. Selain

itu, diadakan juga analisis performa model yang sudah ada beserta *error analysis* untuk mencari hal-hal apa saja yang dapat diperbaiki dalam metodologi ini untuk memperbaiki performa maksimal.

4. Kesimpulan

   Pada tahap terakhir, diambil kesimpulan dari evaluasi yang telah dilakukan untuk menjawab rumusan masalah studi ini.

I.6 Sistematika Pembahasan

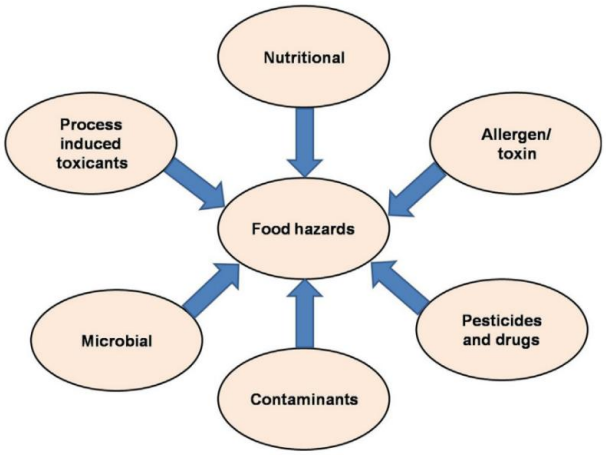
Subbab ini berisi penjelasan ringkas isi per bab. Penjelasan ditulis satu paragraf per bab buku.

BAB II

STUDI LITERATUR

II.1 Bahaya Makanan

Bahaya makanan adalah sebuah agen yang memberi dampak negatif terhadap kesehatan manusia (Singh dkk.., 2019). Bahaya makanan dapat terjadi jika makanan terpapar agen yang berbahaya. Gambar II.1 menunjukkan jenis-jenis bahan yang dapat menyebabkan bahaya makanan.



Gambar II.1. Diagram Jenis Bahaya Makanan

Berdasarkan bentuknya, bahaya makanan dapat diklasifikasikan menjadi fisik, kimia, dan biologis. Selain itu, penipuan makanan (*food fraud*) sera allergen juga merupakan jenis bahaya makanan yang paling sering ditemukan.

II.1.1 Bahaya Fisik (*Foreign Bodies*)

Bahaya fisik (*physical hazard*) adalah benda asing berwujud padat yang pada umumnya tidak terkandung pada makanan atau minuman (*foreign bodies*). Jika terisolasi, bahan ini tidak berbahaya bagi konsumen, namun berisiko merusak

kesehatan tubuh karena kondisinya yang tidak higienis seperti selama proses pengolahan, produksi, penyimpanan, dan distribusi makanan. Bahan asing padat ini tergolong menjadi bahan yang dapat dihindari serta bahan yang tidak dapat dihindari. Contoh bahan asing yang tidak dapat dihindari, yakni bahan yang tidak bisa dipilah secara kasat mata, adalah kotoran pada kentang. Namun sebaliknya, ada beberapa bahan asing yang dapat dihindari, seperti serpihan kaca kecil, potongan perhiasan, dan potongan plastik, yang dapat dicegah dengan metode yang tepat.

II.1.2 Bahaya Kimia

Bahaya kimia (*chemical hazard*) adalah zat asing bersifat molekular yang pada umumnya terserap ke dalam makanan/minuman secara disengaja maupun tidak disengaja. Terdapat dua macam bahaya kimia, yakni zat aditif dan residu pertanian.

Zat aditif makanan adalah zat apa pun yang secara langsung atau tidak langsung menjadi komponen makanan atau mempengaruhi karakteristik makanan. Beberapa contoh zat aditif yang sering tergolong sebagai bahaya makanan mencakup pemanis, pengawet, dan antioksidan.

Zat kimia yang membahayakan tubuh juga dapat berasal dari residu pertanian (*agricultural residues*), seperti pestisida, fungisida, dan herbisida.

II.1.3 Bahaya Biologis

Bahaya biologis (*biological hazard*) disebabkan oleh makhluk hidup yang terdiri atas toksikan mikroba, hewan, dan tumbuhan.

II.1.3.1 Toksikan Mikroba

Bakteri patogen, jamur, parasit, dan virus adalah penyebab penyakit bawaan makanan yang disebabkan oleh mikroba. Penyakit bawaan makanan yang disebabkan oleh bakteri patogen itu sendiri dikenal sebagai infeksi, sementara penyakit akibat produk toksik dari patogen dikenal sebagai keracunan.

- Toksikan Bakteri: Bakteri adalah organisme hidup bersel tunggal yang dianggap sebagai agen penyebab terpenting penyakit bawaan makanan. Makanan yang biasanya mendukung pertumbuhan bakteri adalah susu, telur, unggas, ikan, dan daging. Contoh bakteri yang memproduksi zat beracun yaitu *Bacillus cereus*, *Clostridium botulinum*, dan *Salmonella*.
- Toksikan Jamur: Jamur atau kapang dapat menghasilkan berbagai senyawa kimia (metabolit jamur) yang bersifat aktif secara biologis. Beberapa metabolit jamur sangat diperlukan dalam produksi makanan seperti keju dan obat-obatan (antibiotik). Contoh toksikan jamur yaitu Aflatoksin dan Ochratoxin A.
- Toksikan Virus: Virus adalah parasit intraseluler dan bisa ada dalam makanan tanpa berkembang biak. Contoh toksikan virus yaitu Virus Hepatitis A dan Virus Hepatitis E.
- Protozoa: Protozoa adalah hewan bersel tunggal yang dapat menyebabkan kerusakan organ manusia seperti organ pencernaan jika dikonsumsi. Contoh makhluk hidup dalam kelas Protozoa yaitu cacing gelang (Nematoda) dan cacing pita (Cestoda).

II.1.3.2 Toksikan Hewan

Hewan juga tidak lepas dari zat beracun yang berbahaya bagi konsumen. Salah satu jenis hewan yang memiliki risiko bahaya makanan adalah organisme laut. Walaupun organisme laut merupakan sumber protein hewani yang penting secara global, terdapat banyak spesies organisme laut yang beracun jika tidak diolah dengan baik, seperti kerang (memiliki toksin saxitoxin dan brevetoxin) serta bintang laut (memiliki toksin tetrodotoxin).

II.1.3.3 Toksikan Tumbuhan

Terkadang, makanan memiliki banyak komponen tumbuhan yang mampu mengancam kesehatan tubuh kita terlepas dari komponen tumbuhan baik yang dominan (protein, lemak, dan karbohidrat) maupun tidak dominan (vitamin,

mineral). Komponen berbahaya yang dimaksud yaitu Glucosinolate dalam brokoli, Ptaquiloside dalam pakis, dan lain-lain.

II.1.4 Alergen Makanan

Alergi makanan adalah penyakit kronis yang mengancam jiwa dan sangat membatasi kualitas hidup individu yang mengidapnya. (Greenhawt, 2016). Alergi makanan didefinisikan sebagai jenis reaksi buruk di mana sistem imun berperan. Beberapa makanan alergen yaitu gandum, telur, dan susu.

II.1.5 Penipuan Makanan

Penipuan makanan adalah istilah umum yang mencakup tindakan substitusi, penambahan, pemalsuan, atau kesalahan representasi pada makanan, bahan makanan, atau kemasannya, serta pernyataan yang menyesatkan untuk keuntungan ekonomi. (Spink dan Moyer 2011). Beberapa contoh dari penipuan makanan meliputi (GAO, 2009):

- Penambahan yang berlebihan, yaitu menambahkan es atau air lebih banyak dari yang diizinkan oleh peraturan. Hal ini dilakukan untuk meningkatkan keuntungan dengan menambah bobot melalui es. Air yang ditambahkan dapat mengandung patogen atau bahan kimia (misalnya jika es dibuat dari air kolam).
- Penggantian spesies, yaitu mengganti spesies yang lebih murah dan mengklaimnya sebagai spesies yang lebih mahal. Tindakan ini dilakukan untuk mendapatkan keuntungan dari perbedaan harga. Spesies yang salah label bisa saja beracun atau menyebabkan reaksi alergi.
- Pemalsuan label, yaitu penandaan yang tidak benar pada asal negara, bahan, dan lainnya. Hal ini dilakukan untuk menghindari biaya tambahan dan memaksimalkan keuntungan. Pemalsuan ini bisa menyebabkan alergen yang tidak tertera dan bobot tambahan yang dihasilkan dari bahan lain yang tidak diketahui.

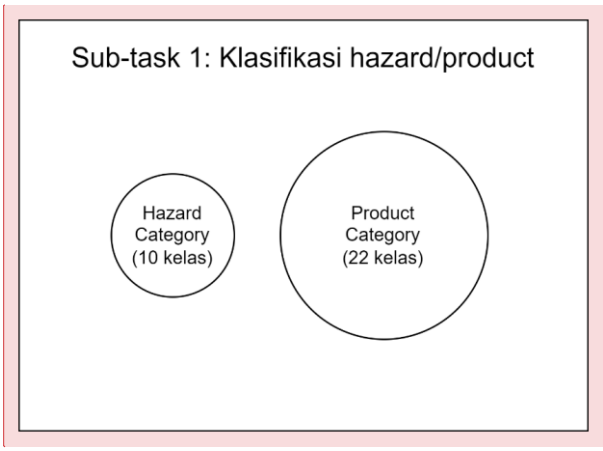
II.2 Analisis Data Studi

Data bahaya makanan yang akan digunakan dalam studi ini terdiri atas data *training*, validasi, dan pengujian. Data yang disediakan oleh SemEval memiliki format *file Comma-Separated Values* (.csv). Terdapat 5.082 baris data dan 11 kolom yang unik dalam data *training*. Gambar II.2. menunjukkan potongan sampel data dari data *training*.

	id	year	month	day	country	title	text	hazard-category	product-category	hazard	product
1323	1323	2015	7	10	us	Gourmet Culinary Solutions Recalls Turkey Sausage Products Due To Possible Foreign Matter Contamination	WASHINGTON, July 10, 2015 – Gourmet Culinary Solutions, a Statham, Ga., establishment, is recalling approximately 495 pounds of turkey sausage products that are part of a frozen entrée that also contains French toast sticks and peaches. The entrées may be contaminated with foreign materials, the U.S. Department of Agriculture's Food Safety and Inspection Service (FSIS) announced today.\n\nThe entrées were produced on May 14, 2015, and the following products are subject to recall: [View Label]\n\n8.25 oz. compartment trays of "Golden Gourmet French Toast Sticks with Turkey Patty & Peaches" with "Use by Date: 11/14/16."\nThe products subject to recall may contain pieces of a conveyor belt inside the packaging. The packages bear establishment number "P-21200" inside the USDA mark of inspection. Individual entrées were distributed to older adults in Georgia as part of the Meals on Wheels program.\n\nThe problem was discovered in 10-lb. bulk packages of the French toast sticks by a customer of the ingredient manufacturer. The customer contacted Gourmet Culinary Solutions. Gourmet Culinary Solutions notified FSIS of the problem, and then began a market withdrawal of the products.\n\nFSIS and the company have received no reports of adverse reactions due to consumption of these products. Anyone concerned about an injury or illness should contact a healthcare provider.\n\nFSIS routinely conducts recall effectiveness checks to verify recalling firms notify their customers of the recall and that steps are taken to make certain that the product is no longer available to consumers.\n\nConsumers or media with questions about the recall can contact Brian Zulaica, director of Gourmet Culinary Solutions, at (770) 725-4620.\n\nConsumers with food safety questions can "Ask Karen," the FSIS virtual representative available 24 hours a day at AskKaren.gov or via smartphone at m.askkaren.gov.The toll-free USDA Meat and Poultry Hotline 1-888-MPHotline (1-888-674-6854) is available in English and Spanish and can be reached from 10 a.m. to 4 p.m. (Eastern Time) Monday through Friday. Recorded food safety messages are available 24 hours a day. The online Electronic Consumer Complaint Monitoring System can be accessed 24 hours a day at: http://www.fsis.usda.gov/reportproblem.\n\n Product Label	foreign bodies	meat, egg and dairy products	plastic fragment	turkey and turkey preparations

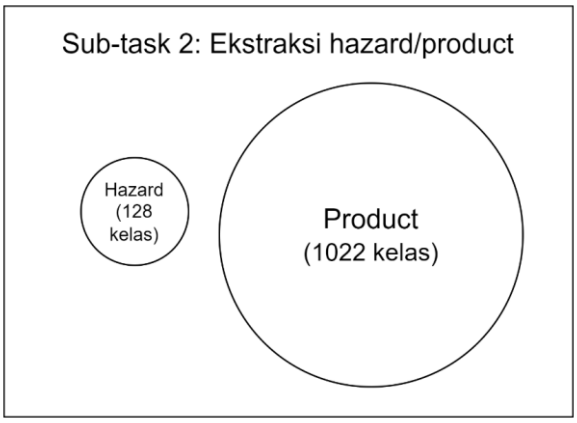
Gambar II.2. Potongan Data Training Food Hazard

Terdapat kolom-kolom seperti "id", "year", "month", "day", dan "country", namun yang paling penting adalah "title", "text", "hazard-category", "product-category", "hazard", dan "product". Kolom "title" menandakan judul dari laporan kontaminasi hazard, sedangkan kolom "text" berupa deskripsi dari bahaya makanan tersebut. Dari "title" dan "text", akan ditentukan klasifikasi kategori bahaya makanan dan product-nya (sub-*task* 1), serta mengekstraksi teks bahaya dan product eksaknya (sub-*task* 2). Kolom "hazard-category" mengandung kelas-kelas bahaya makanan dengan konteks yang lebih luas, misal "foreign bodies", sedangkan kolom "hazard" memiliki makna yang lebih spesifik dan termasuk dalam kolom "hazard-category" pada baris data yang sama, misal "plastic fragment". Perbandingan yang sama juga berlaku untuk "product-category" dan "product".



Gambar II.3. Diagram Ilustrasi Sub-task 1

Commented [MR2]: Kasih contoh kelas

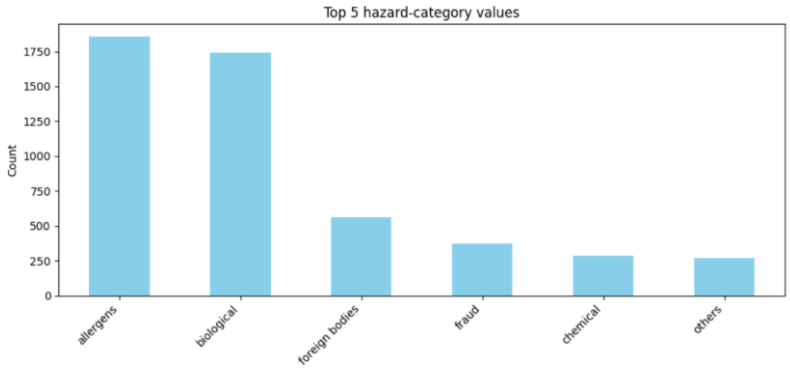


Gambar II.4. Diagram Ilustrasi Sub-task 2

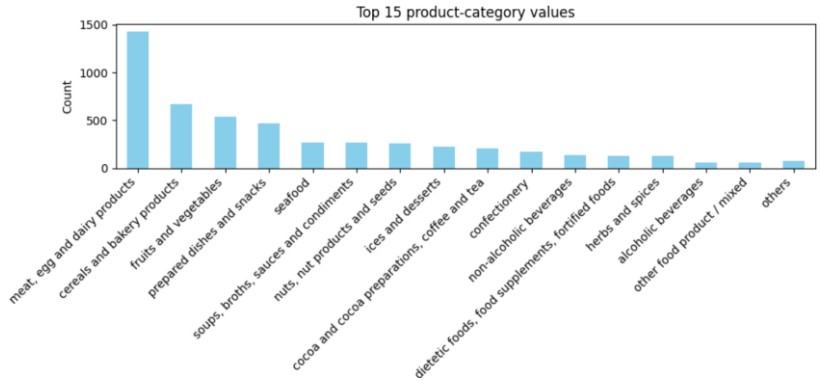
Berikut adalah banyak data dan distribusi data tiap kolom.

id	5082
year	29
month	12
day	31
country	9
title	4948
text	5053
hazard-category	10
product-category	22
hazard	128
product	1022

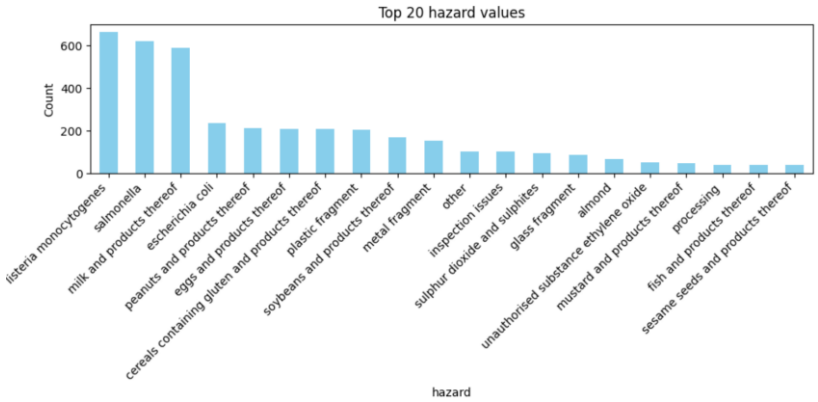
Gambar II.5. Banyak Data Tiap Kolom Data Training



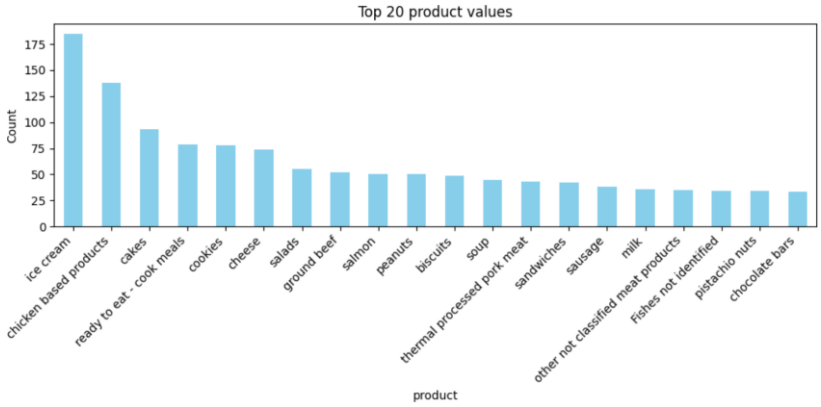
Gambar II.6. Distribusi Data Kolom “hazard-category”



Gambar II.7. Distribusi Data Kolom “product-category”



Gambar II.8. Distribusi Data Kolom “hazard”



Gambar II.9. Distribusi Data Kolom “product”

Terjadi kasus imbalanced data di kolom:

- “hazard-category”, di mana “allergens” dan “biological” menjadi mayoritas data yang paling banyak muncul;
- “product-category”, di mana “meat, egg, and dairy products” menjadi mayoritas data yang paling banyak muncul, dan
- “hazard”, di mana “listeria monocytoneges”, “salmonella”, dan “milk and products thereof” menjadi mayoritas data yang paling banyak muncul.

Terlihat hubungan antara data mayoritas di “hazard-category” dan “hazard” karena bahaya seperti “listeria monocytoneges”, “salmonella” termasuk dalam kategori “biological” dan “milk and products thereof” termasuk dalam kategori “allergens”. Terdapat kemungkinan yang cukup besar data yang bahayanya dilabel sebagai “allergens” dan “biological” memiliki bahaya seperti contoh tersebut.

Untuk melihat lebih lanjut relasi antara deskripsi bahaya (teks dipangkas demi kemudahan membaca) dengan kolom target, tabel II.1 menunjukkan sampel data dengan masing-masing kategori bahaya yang berbeda.

Tabel II.1. Sampel Data dengan “hazard-category” berbeda

title	**text**	**hazard-category**	**product-category**	**hazard**	**product**
Various brands of debittered brewer's yeast recalled due to undeclared peanut	Food Recall Warning (Allergen) - Various brands of debittered brewer's yeast recalled due to undeclared peanut Recall date: February 22, 2019 Reason for recall: Allergen - Peanut	allergens	dietetic foods, food supplements, fortified foods	peanuts and products thereof	brewer's yeast
Cebu's Dried Fish brand Dried Silver Fish (Bol S Dilis) recalled due to histamine - Recalls, advisories and safety alerts – Canada.ca	Notification Cebu's Dried Fish brand Dried Silver Fish (Bol S Dilis) recalled due to histamine Brand(s) Cebu's Dried Fish Last updated 2022-05-20 Summary Product Dried Silver Fish (Bol S Dilis) Issue Food - Chemical	chemical	seafood	toxin	Fishes not identified

Hormel Foods Australia Pty Ltd—Kid's Kitchen—Mini Beef Ravioli in Tomato Sauce, 213g	What are the defects? Incorrect labelling - undeclared allergen - Egg white. What are the hazards? Incorrect labelling	fraud	prepared dishes and snacks	labelling / misdescription	ravioli
Taylor Farms Issues Recall of Products Containing Onions Because of Possible Health Risk	These recalls are due to concerns of the potential for contamination by Salmonella spp. Salmonella is an organism...	biological	fruits and vegertables	salmonella	onions

Berikut adalah analisis korelasi untuk klasifikasi dan ekstraksi teks untuk masing-masing baris data.

a. Data pertama:

- "Hazard category" dikategorikan dalam alergen karena terdapat kata "Allergen" di teks
- "Product category" dikategorikan dalam "dietetic foods, food supplements, fortified foods" karena makanan berupa "debittered brewer's yeast" atau ragi yang merupakan "food supplement"
- "Hazard" dikategorikan dalam "peanuts" karena terdapat frasa "Reason for recall: Allergen - Peanut Hazard" di teks
- "Product" dikategorikan dalam "brewer's yeast" karena makanan berupa "debittered brewer's yeast".

b. Data kedua:

- "Hazard category" dikategorikan dalam *chemical* karena terdapat frasa "Issue Food - Chemical" di teks
- "Product category" dikategorikan dalam "seafood" karena makanan berupa "dried fish"

- “Hazard” dikategorikan dalam “toxin” karena terkontaminasi zat histamine yang tidak tergolong ke dalam golongan “Hazard” lain.
- “Product” dikategorikan dalam “Fishes not identified” karena makanan “dried fish” tidak spesifik terhadap golongan “Product” ikan yang lain.

c. Data ketiga:

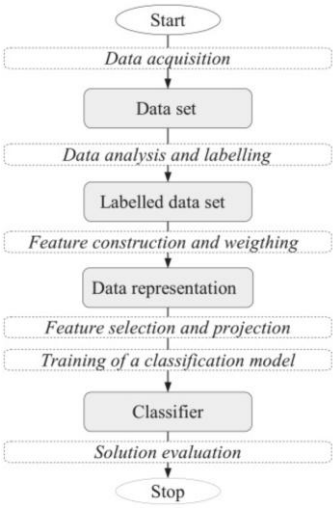
- “Hazard category” dikategorikan dalam *fraud* karena terdapat kata “Incorrect labelling” di teks yang termasuk kesalahan penipuan.
- “Product category” dikategorikan dalam “prepared dishes and snacks” karena ada deskripsi “microwave”.
- “Hazard” dikategorikan dalam “labelling/misdescription” karena terdapat kata “Incorrect labelling” berupa “undeclared allergen - Egg white”.
- “Product” dikategorikan dalam “ravioli” karena makanan berupa “mini beef ravioli”.

d. Data keempat:

- “Hazard category” dikategorikan dalam biological karena terdapat kata “salmonella” dan “organism” di teks yang menunjukkan bahwa makhluk hidup adalah ulah dari bahaya ini.
- “Product category” dikategorikan dalam “fruits and vegetables” karena produk berupa bawang yang termasuk sayuran.
- “Hazard” dikategorikan dalam “salmonella” karena terdapat kata “salmonella” di teks yang di-*emphasize*.
- “Product” dikategorikan dalam “onions” karena terdapat kata “onion” di teks.

II.3 Klasifikasi Teks

Klasifikasi teks adalah masalah konstruksi model yang dapat mengklasifikasikan dokumen baru ke dalam kelas yang telah ditentukan (Liu, 2006; Manning, Raghavan, Schütze, 2008). Klasifikais teks memiliki proses seperti pra-pemrosesan data, transformasi, dan pengurangan dimensi. Kerangka lengkapnya dapat dilihat pada gambar II.10 (Mirończuk & Protasiewicz, 2018).



Gambar II.10. Diagram Kerangka Proses Klasifikasi Teks

Pertama, data dari berbagai sumber, misalnya dari internet atau basis data dikumpulkan dalam format *file* tertentu menjadi sebuah dataset. Selanjutnya, dataset ini di-*preprocess* untuk menghasilkan representasi yang dibutuhkan tergantung metode pembelajaran yang dipilih. Setelah itu, fitur dikonstruksi dari data dan diberi bobot dengan algoritma yang dipilih untuk menghasilkan representasi data yang sesuai. Sesudahnya, jumlah fitur dikurangi menggunakan metode seleksi fitur agar diperoleh representasi fitur data yang paling optimal untuk dilatih.

Model *training* tertentu digunakan untuk melatih fungsi klasifikasi yang akan mengenali konsep target. Ketika model klasifikasi dikembangkan dengan baik, model tersebut dapat mengklasifikasikan data baru. Model yang melakukan klasifikasi menghasilkan keputusan yang mendefinisikan kelas dari setiap vektor input. Terakhir, dilakukan proses evaluasi untuk memperkirakan performa proses klasifikasi teks.

II.3.1 *Imbalanced Classification*

Sebuah dataset dikatakan tidak seimbang jika salah satu kelas dalam dataset tersebut memiliki jumlah sampel yang jauh lebih sedikit dibandingkan dengan kelas-kelas lainnya (Zou dkk., 2016). Secara umum, kelas minoritas dianggap sebagai kelas positif, sedangkan kelas mayoritas dianggap sebagai kelas negatif. Dalam situasi ini, *classifier* mungkin memiliki akurasi yang baik pada kelas mayoritas, tetapi sangat buruk pada kelas minoritas karena pengaruh kelas mayoritas yang lebih besar terhadap kriteria pelatihan tradisional. (Ganganwar, 2012). Berikut adalah beberapa alasan utama mengapa kebanyakan algoritma pembelajaran tidak memiliki performa yang bagus terhadap dataset yang tidak seimbang (Veni & Rani, 2011):

a. Algoritma memprioritaskan akurasi data, sehingga kontribusi kelas minoritas terhadap skor akurasi sangat sedikit.
b. Algoritma mengasumsikan bahwa distribusi data antar kelas mirip.
c. Algoritma mengasumsikan bahwa bahwa error dari kelas-kelas memiliki bobot yang sama.

Pada data yang digunakan, buruknya performa akibat dataset yang tidak seimbang dapat diselesaikan dengan teknik *random oversampling*, *random undersampling*, *directed oversampling*, *directed undersampling*, serta kombinasi dari teknik-teknik tersebut (Chawla, Japkowicz, lcz, 2004).

Oversampling adalah metode paling sederhana untuk meningkatkan jumlah kelas minoritas. *Synthetic Minority Over-sampling Technique* (SMOTE) merupakan pengembangan dari metode *oversampling*, di mana jumlah frekuensi kelas minoritas dilebihkan dengan membuat contoh yang disintesis, dibandingkan *oversampling* dengan pengulangan (Chawla dkk., 2002).

Adapun *undersampling* adalah metode yang menggunakan subset dari kelas mayoritas untuk melatih classifier. Teknik preprocessing yang paling umum adalah *random majority under-sampling* (RUS). Dengan RUS, dataset pelatihan menjadi

lebih seimbang dan proses pelatihan menjadi lebih cepat karena banyak kelas mayoritas diabaikan.

II.4 Ekstraksi Informasi dari Teks dan Deteksi Entitas

Ekstraksi informasi dari teks adalah proses menemukan informasi terstruktur dari teks yang tidak terstruktur atau semi-terstruktur (Jiang, 2012). Ada dua tugas utama dalam ekstraksi informasi: *named entity recognition* (NER) dan ekstraksi relasi. Namun, fokus dalam studi ini adalah pada *named entity recognition*.

II.4.1 Named Entity Recognition (NER)

Commented [MR3]: Add a short description about NER

Named entity adalah rangkaian kata yang menunjukkan entitas di dunia nyata, seperti “Bandung,” “Institut Teknologi Bandung,” dan “Farrel Danendra.” Tugas dari *named entity recognition* (NER) adalah mengidentifikasi sekumpulan *named entity* dalam sebuah teks dan menggolongkannya ke dalam beberapa tipe entitas seperti orang, organisasi, dan lokasi. Terdapat dua pendekatan umum dalam menyelesaikan NER:

a. *Rule-based method*: Sekumpulan aturan ditentukan secara manual atau dipelajari secara otomatis. Setiap token dalam teks direpresentasikan oleh sekumpulan fitur. Teks kemudian dibandingkan dengan aturan-aturan tersebut, dan aturan akan aktif jika ditemukan kecocokan.
b. *Statistic-based learning method*: Dari urutan observasi *feature vector* $x = (x_1, x_2,...,x_n)$ ditetapkan label y_i pada setiap observasi x_i. Dalam proses *sequential labeling*, diasumsikan bahwa label y_i tidak hanya bergantung pada observasi yang bersesuaian x_i tetapi juga pada observasi dan label lain dalam urutan tersebut. Ada banyak algoritma yang cocok digunakan dalam metode ini, seperti *Hidden Markov Models*, *Maximum Entropy Markov Models*, dan *Conditional Random Fields* (CRF).

NER merupakan langkah yang sangat krusial dalam proses deteksi entitas pada studi ini. Deteksi entitas mengacu pada proses mengidentifikasi dan mengekstraksi entitas tertentu (dalam hal ini, produk dan bahaya) dari data teks.

Proses NER dapat dilakukan dengan mengimplementasikan model yang telah di-*fine tune* oleh dengan dataset objek tertentu demi kepentingan spesialisasi deteksi entitas jenis spesifik. Salah satu model NER yang dimaksud adalah "Dizex/InstaFoodRoBERTa-NER" pada Huggingface yang mampu mengekstraksi segala entitas makanan, baik dalam kata singular maupun dalam bentuk frasa. Model ini telah di-*pretrain* dengan dataset yang terdiri atas 400 postingan Instagram yang berkaitan dengan makanan. Proses NER yang dilakukan oleh model ini telah terbilang sangat baik, dengan nilai *precision* 0.89, *recall* 0.93, dan *f1-score* 0.91.

II.5 Augmentasi Data

Strategi augmentasi data (*Data Augmentation*) digunakan untuk meningkatkan variasi data pelatihan tanpa harus mengumpulkan data baru secara eksplisit. Sebagian besar strategi ini menambahkan salinan data yang telah dimodifikasi sedikit atau membuat data sintetis. Tujuannya adalah agar data yang diperluas ini dapat bertindak sebagai regularisasi dan mengurangi overfitting saat melatih model pembelajaran mesin (Shorten & Khoshgoftaar, 2019; Hernández-García & König, 2020). Dalam bidang NLP, augmentasi data semakin diminati, terutama karena semakin banyak penelitian di domain dengan sumber daya terbatas, munculnya tugas-tugas baru, serta meningkatnya popularitas jaringan saraf skala besar yang membutuhkan banyak data pelatihan (Feng dkk., 2021). Pustaka NLPAug merupakan peralatan yang sangat umum digunakan dalam proses augmentasi data.

Ada beberapa metode untuk melakukan augmentasi data teks pada tingkat kata. Salah satu metode yang paling umum adalah transformasi teks melalui parafrase, di mana kata-kata tertentu dalam teks diganti dengan sinonim. Penggantian sinonim ini merupakan pendekatan yang alami dan efektif untuk memperkaya data teks (Wei & Zou, 2019). Biasanya, metode ini memilih kata dalam suatu kalimat asli dan menggantinya dengan kata lain yang memiliki makna serupa. Salah satu penerapan pertama dari metode ini dalam augmentasi data adalah penggantian ekspresi temporal dengan sinonim potensial yang diambil dari WordNet (Miller dkk., 1990).

Metode lain dalam augmentasi teks adalah penggantian berbasis *embedding* kontekstual. Pendekatan ini mencari kata-kata yang paling sesuai dengan konteks kalimat tanpa mengubah makna aslinya. Untuk melakukannya, setiap kata dalam teks dikonversi ke dalam ruang representasi laten, di mana kata-kata dengan konteks serupa memiliki posisi yang berdekatan. Pemilihan kata dalam metode ini mengikuti hipotesis semantik distribusional (Firth, 1962), yang menyatakan bahwa kata-kata dengan makna serupa muncul dalam konteks yang mirip. Keunggulan metode ini dibandingkan penggantian sinonim adalah tidak terbatas pada basis data seperti WordNet dan mampu menghasilkan kalimat yang lebih gramatikal serta alami (Aroyehun & Gelbukh, 2018).

II.6 *Large Language Model*

Commented [MR4]: Add short descriptions of roberta and deberta and other models experimented

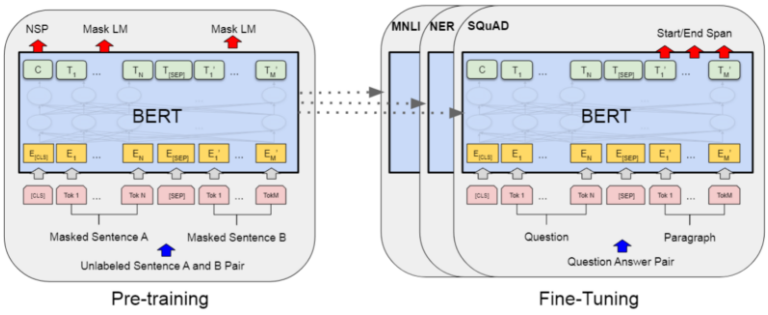
Hari ini, sebagian besar solusi NLP (*Natural Language Processing*) berbasis pada model *deep learning* yang diimplementasikan menggunakan arsitektur jaringan saraf. Baru-baru ini, arsitektur *transformer* seperti BERT yang diimplementasikan dengan mekanisme *self-attention* telah menjadi *state-of-the-art*. Di domain umum, model NLP berbasis *transformer* telah mencapai kinerja terbaik untuk *named entity recognition* dan ekstraksi relasi. Biasanya, transformer dilatih dalam dua tahap: *pretraining* model dan *fine-tuning*. Satu model yang telah dilatih sebelumnya dapat diterapkan untuk menyelesaikan banyak tugas NLP melalui *fine-tuning*. Model ini dikenal sebagai *transfer learning* (Bommasani dkk., 2022). Sebuah studi menunjukkan bahwa model *transformer* yang dilatih menggunakan data teks yang besar memiliki performa yang jauh lebih baik daripada model NLP sebelumnya (Yang dkk., 2022).

II.6.1 *Bidirectional Encoder Representations from Transformers* (BERT)

BERT (*Bidirectional Encoder Representations from Transformers*) adalah model *deep learning* yang dirancang untuk melatih representasi dari data teks yang tidak diberi label (Devlin dkk., 2019). Hal yang membedakan BERT dari model lainnya yaitu pada metodologi representasi teksnya dan paradigma pelatihannya.

Pelatihannya terdiri dari dua fase yang berbeda: *pre-training* pada korpus teks yang belum diberi label dan *fine-tuning* pada dataset yang telah diberi label (Vaswani, 2017). *Fine-tuning* atau bahkan *re-training* dapat dilakukan pada tugas dan dataset yang berbeda (Sun dkk., 2019).

Gambar II.11 adalah gambar arsitektur BERT untuk proses *pre-training* dan *fine-tuning* (Devlin dkk., 2019).



Gambar II.11. Arsitektur BERT pre-training dan fine-tuning

Ciri khas BERT adalah arsitekturnya yang homogen di berbagai tugas. Arsitektur yang digunakan dalam *pre-training* sama dengan *fine-tuning*, kecuali dalam lapisan output. Parameter model yang di-*pre-train* digunakan untuk menginisialisasi model *fine-tuning*.

Bentuk arsitektur model BERT adalah *encoder Transformer* dua arah yang memiliki beberapa lapisan. Misalnya jumlah lapisan L, ukuran tersembunyi H, dan jumlah *self-attention head* A. Terdapat dua ukuran model BERT yang sering digunakan, yaitu BERTBASE (L=12, H=768, A=12, Total Parameter=110M) dan BERTLARGE (L=24, H=1024, A=16, Total Parameter=340M). Untuk membuat BERT dapat menangani berbagai tugas *downstream*, representasi input BERT mampu merepresentasikan baik satu kalimat maupun sepasang kalimat dalam satu urutan token. "Kalimat" dapat berupa teks kontinu acak, bukan hanya kalimat linguistik. "*Sequence*" merujuk pada urutan token input ke BERT, yang bisa berupa satu kalimat atau dua kalimat yang dikemas bersama.

[CLS] adalah simbol khusus yang ditambahkan di depan setiap contoh input, sedangkan [SEP] adalah token pemisah khusus (misalnya, untuk memisahkan pertanyaan/jawaban). *Hidden state* akhir yang sesuai dengan token ini digunakan sebagai representasi agregat dari urutan untuk tugas klasifikasi. Selain dipisahkan dengan token [SEP], kalimat dapat dibedakan dengan menambahkan *embedding* yang dipelajari ke setiap token untuk menunjukkan apakah token tersebut berasal dari kalimat A atau kalimat B. Pada Gambar II.11, *embedding input* direpresentasikan sebagai E, vektor tersembunyi akhir dari [CLS] sebagai C, dan vektor tersembunyi akhir untuk token input ke-i sebagai Ti.

II.6.1.1 Tensor dan Formatnya

Dalam pembelajaran mesin, pemrosesan batch (*batch processing*) adalah teknik yang memproses data dalam jumlah besar secara berkelompok (*batch*) yang telah ditentukan sebelumnya, daripada memproses data secara langsung atau terus-menerus. Dalam pemrosesan batch, setiap batch dikonversi menjadi token menggunakan tokenizer BERT, yang menghasilkan input_ids, attention_mask, dan labels untuk setiap urutan yang telah ditokenisasi. Informasi ini kemudian digunakan untuk menyempurnakan model (*fine-tuning*) (RajkumarDheivanayahi dkk, 2024).

a. input_ids: Merupakan ID dari urutan input setelah melalui proses tokenisasi.
b. attention_mask: Masker biner yang dihasilkan oleh *tokenizer*, menunjukkan token mana yang perlu diperhatikan dan mana yang tidak, dalam bentuk daftar angka 1 dan 0 (Hussain dkk, 2024).
c. labels: Menyimpan label sebenarnya dari dataset. Digunakan untuk membandingkan prediksi model dengan nilai sebenarnya guna mengevaluasi performa model (Alshawabkeh, 2024).

Untuk mempermudah pemrosesan data dalam pelatihan dan validasi model, DataLoader dari PyTorch dibuat menggunakan data input (x_inputs, x_masks) dan label yang sesuai (y_labels). DataLoader memungkinkan data dibagi menjadi *batch* yang lebih kecil dan diproses secara paralel, meningkatkan efisiensi pelatihan.

Tensors PyTorch, y_train_labels dan y_val_labels, dibuat dari label pelatihan (y_train) dan label validasi (y_val). Tensor ini digunakan sebagai input ke dalam DataLoader. Hasil akhirnya adalah dua DataLoader: train_dataloader untuk data pelatihan dan val_dataloader untuk data validasi, yang keduanya siap digunakan dalam pelatihan model (Rubio, Almeida, & Seguda-Bedmar, 2023).

Sebelum memulai pelatihan, diperlukan data collator, yang bertugas mengambil sampel dari dataset dan menyusunnya ke dalam batch. Hasilnya berupa objek berbentuk kamus (*dictionary*). Beberapa parameter penting dalam menginisialisasi *data collator* adalah MLM (*Masked Language Model*) yang memproses sampel dalam *batch* dan mlm_probability yakni persentase token yang akan dimasker selama proses *pre-processing* (Rothman, 2022).

Selanjutnya, input_ids dan attention_mask yang telah dikumpulkan oleh *data collator* diteruskan ke model untuk memperoleh output. Output ini berisi last_hidden_state, yaitu urutan status tersembunyi (*hidden states*) dari lapisan terakhir model. last_hidden_state mengandung representasi vektor dari semua token dalam paket, termasuk token khusus seperti CLS dan SEP.

II.6.1.2 Parameter Pelatihan Model dan Optimisasi BERT

Terdapat banyak parameter yang dapat digunakan untuk mendukung dan meningkatkan kinerja pelatihan model BERT, di antaranya epoch dan *optimizer*.

Epoch adalah jumlah pengulangan proses pembelajaran yang dilakukan dalam pembelajaran mesin (Hastomo dkk, 2021). Satu epoch berarti model telah memproses seluruh sampel dalam dataset pelatihan dan memperbarui parameter berdasarkan nilai *loss* yang dihitung (Malahina dkk, 2024). Semakin besar jumlah epoch, semakin lama waktu pemrosesan yang dibutuhkan. Namun, peningkatan jumlah epoch tidak selalu menjamin akurasi yang lebih tinggi.

Optimizer digunakan untuk menentukan sejauh mana bobot dan *learning rate* harus diubah guna mengurangi nilai loss (Llugsi dkk, 2021). Salah satu *optimizer* yang paling dikenal adalah Adam (*Adaptive Moment Estimation*), yang mempertahankan

kestabilan pembaruan parameter terhadap perubahan skala gradien. Versi variannya, AdamW, menambahkan regularisasi berbasis weight decay dalam proses optimasi. Dalam AdamW, *weight decay* hanya dilakukan setelah ukuran langkah (*step size*) dari setiap parameter dikendalikan.

II.6.1.3 BERT *Extension* Menggunakan *Embedding* dan Duplikasi Token CLS

Model BERT menggunakan token khusus untuk menangani berbagai tugas serta struktur input yang diperlukan. Salah satu token tersebut adalah token [CLS], yang berfungsi untuk merangkum seluruh urutan input (Kraišniković dkk, 2025). Representasi yang terkait dengan token ini sering digunakan dalam tugas klasifikasi atau tugas lain yang memerlukan pemahaman terhadap keseluruhan konteks suatu teks. Token [CLS] digunakan dalam klasifikasi serta analisis *embedding*.

Menggandakan *embedding* token [CLS] bertujuan untuk meningkatkan dimensi fitur dalam representasi urutan, sehingga dapat memberikan lebih banyak informasi kepada model klasifikasi (Behera & Dash, 2022). Pendekatan ini dapat membantu dalam meningkatkan pemisahan antar kelas (class separability). Alih-alih hanya menggunakan satu *embedding* [CLS], metode ini menciptakan versi yang diperluas untuk menangkap lebih banyak fitur dari teks yang dianalisis.

II.6.1.4 Model-Model Varian BERT

Terdapat beberapa varian model BERT yang di-*enhance* dalam beberapa aspek, seperti RoBERTa dan DeBERTa.

RoBERTa (*Robustly optimized BERT approach*) mengoptimalkan berbagai hiperparameter utama serta ukuran data pelatihan untuk meningkatkan performa model (Liu dkk, 2019). RoBERTa dikembangkan untuk memperbaiki keterbatasan BERT, yang sebelumnya mengalami pelatihan yang kurang optimal. Model ini dilatih dengan beberapa perubahan signifikan, seperti *dynamic masking*, penggunaan kalimat penuh tanpa *Next Sentence Prediction* (NSP) *loss*, *mini-batch*

yang lebih besar, serta *Byte-Pair Encoding* (BPE) berbasis *byte-level* yang lebih luas.

a. NSP (*Next Sentence Prediction*) adalah fungsi *loss* klasifikasi biner yang digunakan untuk memprediksi apakah dua segmen teks dalam dataset aslinya saling berurutan atau tidak.
b. BPE (*Byte-Pair Encoding*) adalah metode representasi teks yang menggabungkan pendekatan berbasis karakter dan kata, sehingga dapat menangani kosakata yang sangat besar dalam korpus bahasa alami (Sennrich dkk., 2016).

Sementara itu, DeBERTa (*Decoding-enhanced BERT with disentangled attention*) meningkatkan model BERT dan RoBERTa dengan dua teknik utama (He dkk, 2020):

a. Mekanisme *Disentangled Attention*: Setiap kata direpresentasikan dengan dua vektor, yaitu vektor konten dan vektor posisi. Bobot perhatian (*attention weights*) antar kata dihitung menggunakan matriks yang mempertimbangkan hubungan antara konten dan posisi relatifnya.
b. *Enhanced Mask Decoder*: Teknik ini digunakan dalam lapisan *decoder* untuk memasukkan informasi posisi absolut dalam prediksi token yang dimasker selama prapelatihan model.

Dengan kombinasi teknik ini, DeBERTa dapat meningkatkan pemahaman konteks dan hubungan antar kata lebih baik dibandingkan pendahulunya.

II.7 Metrik Evaluasi

Salah satu metrik evaluasi yang paling banyak digunakan untuk mengukur performa model *machine learning* adalah akurasi, presisi, dan *recall*. (Sammut & Webb, 2011). Akurasi adalah proporsi dari jumlah prediksi yang benar terhadap total prediksi. Akurasi memiliki rumus:

$$Accuracy = \frac{\sum_{i=1}^{n} N_{ii}}{\sum_{i=1}^{n} \sum_{j=1}^{n} N_{ij}}$$

Presisi adalah ukuran ketepatan model, dengan syarat bahwa model memprediksi suatu kelas tertentu. Presisi memiliki rumus:

$$Precision_i = \frac{N_{ii}}{\sum_{k=1}^{n} N_{ki}}$$

Recall adalah ukuran kemampuan model untuk memilih instansi dari suatu kelas tertentu dalam dataset. *Recall* memiliki rumus:

$$Recall_i = \frac{N_{ii}}{\sum_{k=1}^{n} N_{ik}}$$

Selain itu, metrik evaluasi yang sangat umum dipakai adalah F1-*score*, yaitu rata-rata harmonik dari presisi dan *recall* dengan rumus:

$$F - score_i = \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i}$$

F1-*score*, yang juga dikenal sebagai koefisien kesamaan Dice, merupakan rata-rata harmonis dari presisi dan *recall*, memberikan keseimbangan antara keduanya (Akosa, 2017). Beberapa jenis F1-*score* termasuk *micro* F1-*score*, *macro* F1-*score*, dan *weighted* F1-score. Metrik seperti akurasi serta F1-*score* mikro dan makro direkomendasikan untuk mengevaluasi kinerja pengklasifikasi multilabel (Rainio, Teuho, & Klén, 2024). Literatur terbaru menunjukkan bahwa akurasi dan F1-*score* adalah metrik kinerja yang paling sering digunakan untuk pengklasifikasi multilabel (Heydarian, Doyle, Samavi, 2022).

Macro F1-*score* menghitung F1-*score* secara independen untuk setiap label, lalu mengambil rata-ratanya, memberikan bobot yang sama untuk setiap kelas. *Macro* F1-*score* cocok untuk mengevaluasi model yang mengendalikan data yang tidak seimbang distribusinya karena menghitung semua F1-*score* setiap kelas secara terpisah dan menggabungkannya dalam rata-rata.

Berikut adalah rumus dari *macro* F1-*score*:

$$F1_{macro} = \frac{1}{N} \sum_{i=1}^{N} \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i}$$

dengan N adalah banyak kelas yang ada dalam data.

II.8 Studi Terkait Pemodelan Bahaya Makanan

II.8.1 Studi Terkait Klasifikasi Bahaya Makanan

Salah satu eksperimen klasifikasi bahaya makanan menggunakan pendekatan machine learning dengan Random Forest Model untuk menentukan variabel prediktor utama dalam perubahan perilaku keamanan pangan selama pandemi. Sebuah survei daring dikembangkan untuk mengumpulkan data terkait persepsi risiko konsumen di AS terhadap COVID-19 dan penyakit bawaan makanan (foodborne illness/FBI), praktik perilaku keamanan pangan, serta karakteristik demografis. Survei dilakukan pada sepuluh waktu berbeda dari tahun 2020 hingga 2022, menghasilkan setidaknya 700 data pada setiap waktu pengambilan, dengan total 7.355 data.

Model Random Forest digunakan untuk memprediksi 14 perilaku terkait keamanan pangan, dengan pembagian data 70% untuk pelatihan dan 30% untuk pengujian. Saat membangun model, variabel perilaku keamanan pangan berbentuk biner diseimbangkan menggunakan teknik Synthetic Minority Oversampling Technique (SMOTE). Model untuk memprediksi perilaku cuci tangan sebelum memasak dan cuci tangan setelah makan menunjukkan kinerja yang baik, dengan skor F-1 masing-masing sebesar 0,93 dan 0,88. Variabel terkait sikap (attitudes-related variables) ditemukan sebagai faktor penting dalam memprediksi perilaku keamanan pangan.

II.8.2 Studi Terkait Augmentasi Data Bahaya Makanan

Dalam sebuah penelitian mengenai dampak augmentasi data menggunakan ChatGPT-4o-mini terhadap analisis bahaya pangan dan produk, model FLAN-T5 menunjukkan kinerja terbaik dengan penerapan augmentasi. Setelah jumlah sampel dalam beberapa kelas ditingkatkan melalui augmentasi, performa model mengalami peningkatan yang signifikan (Rasheed dkk., 2025). Nilai *f1-score* model FLAN-T5 sebelum dan sesudah augmentasi data adalah 76.14 dan 78.10.

Dalam klasifikasi cuitan yang mengandung unsur mengganggu, metode embedding kontekstual terbukti efektif (Wang & Yang, 2015). Dengan menggunakan algoritma k-nearest-neighbor untuk mengidentifikasi *embedding* yang paling sesuai sebagai pengganti kata dalam data pelatihan, model mengalami peningkatan F1-score hingga 2,4 poin dibandingkan model dasar.

Pustaka NLPAug memiliki peran penting dalam berbagai penelitian, seperti yang dilakukan dalam studi yang memanfaatkan model AraBERT, MARABERT, dan AraELECTRA (Fsih dkk, 2019) dengan BERT untuk menggantikan *embedding* kontekstual dalam deteksi sarkasme pada tweet berbahasa Arab (Kamr & Mohamed, 2022). Skor F1 yang diperoleh dari tiga model berbeda adalah 87%, 90%, dan 91%. Selanjutnya, model-model ini diperkuat dengan metode *voting* keras dalam model ansambel, yang akhirnya menghasilkan skor F1 sebesar 93%.

II.8.3 Studi Terkait Ekstraksi Informasi Bahaya Makanan

Studi kasus pertama yang meneliti ekstraksi informasi terkait bahaya makanan adalah *fine-tuning* model BERT dalam ekstraksi informasi makanan (Stojanov dkk, 2021). Data korpus yang digunakan berasal dari FoodBase yang terdiri atas 200 resep makanan untuk versi yang dikurasi dan 22.000 resep makanan untuk versi yang tidak dikurasi. Data semantik seperti korpus Hansard dan ontologi FoodOn juga digunakan. Model BERT yang digunakan untuk *fine-tuning* terdiri atas model BERT dan dua model BioBERT. Semua model BERT menghasilkan *macro F1-score* antara 93.30% sampai 94.31% untuk membedakan entitas makanan dan non-makanan, serta *macro F1-score* sebesar 73.39% sampai 78.96% untuk memprediksi label semantic makanan.

Studi kasus kedua mengusulkan kerangka *Named Entity Extraction for Food Safety Monitoring* (NEE-FSM) yang terdiri atas tiga modul, yakni *Heuristic Entity Extraction* (HEE), *Named Entity Recognition* (NER), dan *Named Entity Matching* (NEM) (Lee, 2023).

Modul *Heuristic Entity Extraction* (HEE) berfungsi membangun kamus dasar dari entitas nama yang umum dalam keamanan makanan dari berbagai sumber pengetahuan. Menggunakan *Regular Expression Parser*, frasa entitas makanan diekstraksi dengan menggabungkan tag POS (*Part-of-Speech*) dengan aturan tata bahasa yang disesuaikan secara manual (misalnya, menggabungkan "Adjective (JJ)", "Nouns (NNS)" dan "Verbs (VBN)").

Modul *Named Entity Recognition* (NER) berfungsi menambahkan dan meningkatkan kamus dasar dengan entitas nama baru dari dataset keamanan makanan. Untuk mengkodekan urutan input tingkat kata menjadi token, digunakan *Pre-trained BERT Tokenizer* (*BertTokenizerFast*). Tokenizer BERT menghasilkan urutan token *word-piece* yang diindeks dengan token ID dari kosa kata BERT yang telah dilatih. Model BERT yang telah di-*pre-trained* dilakukan *fine-tuning* untuk tugas NER dalam keamanan makanan dengan menggunakan beberapa dataset patokan. Sebagai hasilnya, model BERT yang di-*fine-tune* untuk NER digunakan untuk mengekstraksi entitas nama terkait keamanan makanan dari dataset keamanan makanan untuk memperkaya kamus dasar.

Modul *Named Entity Matching* (NEM) berfungsi melakukan ekstraksi entitas nama berbasis kamus menggunakan kamus yang telah diperluas. Untuk melakukan ekstraksi NEM, istilah entitas dari setiap kamus keamanan makanan diurutkan berdasarkan panjang kata untuk menghasilkan sekumpulan kamus n-gram. Selanjutnya, dilakukan proses ekstraksi entitas dengan mencocokkan istilah dari setiap kalimat input dengan kamus n-gram.

Berikut adalah tabel perbandingan hasil eksperimen bahaya makanan dari studi literatur yang dilakukan.

Tabel II.2. Perbandingan Studi Bahaya Makanan

Nama Artikel	**Dataset**	**Metode Pelatihan**	**Hasil**

Lee, Z. H. N. (2023). Named entity extraction for food safety events monitoring.	FoodDB dan FoodBase berupa corpus dan jsonl.	Heuristic Entity Extraction (HEE), Named Entity Recognition (NER), dan Named Entity Matching (NEM) → NER menggunakan Pre-trained BERT Tokenizer (BertTokenizerFast).	NER Benchmark Dataset: Task Specific Fine-tuning (TSF) BERT memiliki performa terbaik dengan skor precision, recall, dan F1-score 83.02%, 86.84%, 84.89%.
Rasheed, A. F., Zarkoosh, M., Chasib, S. A., & Abbas, S. F. (2025). Data Augmentation to Improve Large Language Models in Food Hazard and Product Detection.	Data *training* (5082 sampel), validasi (565 sampel), penguji (997 sampel)	Model FLAN-T5 dan augmentasi data menggunakan ChatGPT-4o-mini	*F1-score* model FLAN-T5 sebelum dan sesudah augmentasi data adalah 76.14 dan 78.10.
W. Y. Wang and D. Yang, "That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying	3375 cuitan dengan hashtag #petpeeve	Memakai model SAGE (Eisenstein et al., 2011) dan embedding kontekstual.	Model mengalami peningkatan F1-score hingga 2,4 poin dibandingkan model dasar.

behaviors using #petpeeve tweets," in Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing, 2015, pp. 2557–2563. doi: 10.18653/v1/d15-1306.			
Kamr AM, Mohamed EH (2022) akaBERT at SemEval-2022 Task 6: An Ensemble Transformer-based Model for Arabic Sarcasm Detection. In: SemEval 2022—16th international workshop on semantic evaluation, proceedings of the workshop, pp 885–890.	Dataset berbahasa Arab, terdiri atas 3102 cuitan dengan 2357 cuitan non-sarkastik and 745 cuitan sarkastik. Dataset penguji memiliki 1400 cuitan.	Melakukan *fine-tuning* model AraBERT, MARABERT, dan AraELECTRA, serta augmentasi data menggunakan NLPAug.	Skor F1 yang diperoleh dari tiga model berbeda adalah 87%, 90%, dan 91%. Selanjutnya, model-model ini diperkuat dengan metode *voting* keras dalam model ansambel, yang akhirnya menghasilkan skor F1 sebesar 93%.
Berglund, Z., Kontor-Manu, E., Jacundino, S. B., & Feng, Y. (2024). Random	Survei online sebanyak 10 kali dari 2020 sampai 2022, dengan jumlah	Random forest + 70:30 training testing split + SMOTE	Skor F-1 masing-masing sebesar 0,93 dan 0,88 untuk perilaku cuci tangan sebelum memasak dan cuci

forest models of food safety behavior during the COVID-19 pandemic.	data 7355.		tangan setelah makan
Stojanov R, dkk. (2021). A Fine-Tuned Bidirectional Encoder Representations From Transformers Model for Food Named-Entity Recognition: Algorithm Development and Validation	Korpus FoodBase dan data semantik seperti korpus Hansard dan ontologi FoodOn	Satu model BERT dan dua model BioBERT	*Macro F1-score* antara 93.30% sampai 94.31% untuk membedakan entitas makanan dan non-makanan, serta *macro F1-score* sebesar 73.39% sampai 78.96% untuk memprediksi label semantic makan

BAB III

ANALISIS DAN RANCANGAN SOLUSI

III.1 Analisis Penyelesaian Masalah

Salah satu masalah utama dalam studi ini adalah ketidakseimbangan dataset makanan yang telah disediakan. Dibandingkan kondisi yang diharapkan, data bahaya makanan masih memiliki distribusi kelas yang tidak seimbang, menyebabkan ketidakakuratan model konvensional dalam melakukan prediksi. Hal ini telah terbukti pada subbab II.2, khususnya untuk kolom "hazard-category" dan "product-category".

Masalah lain yang juga dihadapi adalah beberapa jenis model BERT yang dapat dipilih, seperti BERT biasa ("bert-base-uncased"), RoBERTa, dan DeBERTa. Eksperimen sangat penting dilakukan untuk menentukan model apa yang menghasilkan nilai paling optimal untuk prediksi klasifikasi dan ekstraksi data bahaya makanan. Model RoBERTa dan DeBERTa belum tentu memiliki performa yang lebih baik dibandingkan BERT tergantung kompleksitas dataset dan *hyperparameter* yang telah diatur.

Masalah berikutnya adalah menentukan *hyperparameter* yang tepat untuk meraih hasil model maksimal untuk klasifikasi teks dan ekstraksi entitas. *Hyperparameter* yang ditentukan dapat berupa banyak kata yang diaugmentasi atau jumlah *epoch* pelatihan model.

III.2 Rancangan Solusi

III.3 Pengembangan Solusi

BAB IV

<EVALUASI>

Tujuan penulisan bab ini adalah untuk menunjukkan seberapa jauh solusi yang diuraikan pada bagian sebelumnya dapat menyelesaikan permasalahan utama Tugas Akhir. Metode yang dipakai untuk melakukan evaluasi dapat bermacam-macam, bergantung pada jenis permasalahannya.

BAB V

KESIMPULAN DAN SARAN

Bab Kesimpulan dan Saran merupakan penutup dari bagian utama Laporan Tugas Akhir. Fokuskan kesimpulan pada hal-hal baru yang relevan dengan ketercapaian tujuan Tugas Akhir terkait dengan permasalahan yang diselesaikan dalam Tugas Akhir. Saran berisi kajian hal-hal yang masih dapat dikembangkan lebih lanjut.

DAFTAR REFERENSI

Balabanovic, M. (1998). *Learning to surf: Multi-agent systems for adaptive web page recommendation*. Doctoral dissertation, Stanford University, Menlo Park, CA: Department of Computer Science.

McKusick, K.B., & Langley, P. (1991). Constraints on tree structure in concept formation. Prosiding *The 21st ACM-SIGIR International Conference on Research and Development in Information Retrieval*, 206-214. New York, NY:ACM Press.

Mitchell, T.M. (1997). *Machine Learning*. New York, NY: McGraw-Hill.

Pazzani, M., & Billsus, D. (1997). Learning and revising user profiles: The identification of interesting web sites. *Machine Learning, 27*, 313-331.

Lampiran A. Contoh Judul Lampiran

A.1 Contoh Judul Anak Lampiran

Contoh anak lampiran