UNIVERSITI TEKNOLOGI MALAYSIA

MDS Project

BERT-BASED SEMANTIC SIMILARITY OF MALAYSIAN LEGAL PRECEDENTS

NAME: MUHAMMAD HAZIQ BIN MOHAMAD (MCS241036)
SUPERVISED BY: PROF. MADYA DR. MOHD. SHAHIZAN OTHMAN

Innovating Solutions

LINK OF THE VIDEO PRESENTATION

Please click here to open the video presentation.

INTRODUCTION & MOTIVATION

The Challenge of Navigating Malaysian Legal Data:

- **Information Overload:** The Malaysian legal system is experiencing an exponential growth of digital case law, making manual research increasingly difficult.
- **High Cost of Inefficiency:** Manual precedent research is a time-consuming and costly process for legal practitioners, impacting productivity and access to justice.
- **Limitations of Current Tools:** Traditional keyword-based search systems often fail to capture the semantic context of legal arguments, leading to missed precedents and irrelevant results.

Project Goal: To develop and evaluate a deep learning model that understands the meaning behind legal text, providing a more intelligent and accurate method for case discovery.

PROBLEM STATEMENT

The "Vocabulary Mismatch" Problem: A relevant case may discuss the same legal concept using entirely different words or phrasing. Keyword search will fail to connect them.

Inability to Discover Conceptual Links: Practitioners struggle to find cases with similar factual matrices that are described differently.

Wasted Resources: A significant amount of time is spent sifting through irrelevant cases returned by imprecise searches, hindering case preparation and strategy.

Research Question

- RQ: What are the current semantic similarity methods used in legal Natural Language Processing (NLP), and how effective are they in the legal domain?

- RQ: How can a BERT-based model be fine-tuned specifically for Malaysian legal case texts to improve semantic similarity detection?

- RQ: How does the fine-tuned BERT-based model perform compared to traditional and existing transformer-based approaches

Research Objectives

- RO: To investigate existing semantic similarity methods in legal Natural Language Processing (NLP)

- RO: To develop and fine-tune a BERT-based model tailored for Malaysian legal case texts

- RO: To evaluate and visualize model performance using comprehensive metrics and analysis

LITERATURE REVIEW

Category	Method/Model	Description	Limitation/Advantage	References
Traditional Methods	Bag-of-Words (BoW)	Represents text based on word frequency.	Ignores word order and context; suffers from vocabulary mismatch.	Salton, G., & McGill, M. J. (1983)
	TF-IDF	Weighs word frequency by how rare a word is across documents.	Cannot differentiate meaning in context; e.g., "the court dismissed the appeal" vs "the appeal dismissed the court."	Spärck Jones, K. (1972)
Transformer Models	BERT	Reads the entire sentence bidirectionally to understand context from both directions.	Strong contextual understanding, but not specifically tuned for sentence similarity tasks.	Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019)
	Legal BERT, Legal-Longformer, Legal XLNet	Pre-trained on legal texts, improving understanding of legal terminology and syntax.	Tailored for legal domain, offering more accurate results in legal NLP tasks.	1. Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). 2. Beltagy, I., Peters, M. E., & Cohan, A. (2020). 3. Yang, Z., Dai, Z., Yang, Y., Carbonell, J.,
	Sentence-BERT (SBERT)	Modified BERT architecture to generate meaningful sentence embeddings for semantic similarity tasks.	More efficient and effective for comparing sentence meanings than vanilla BERT.	Reimers, N., & Gurevych, I. (2019)

Problem Identification

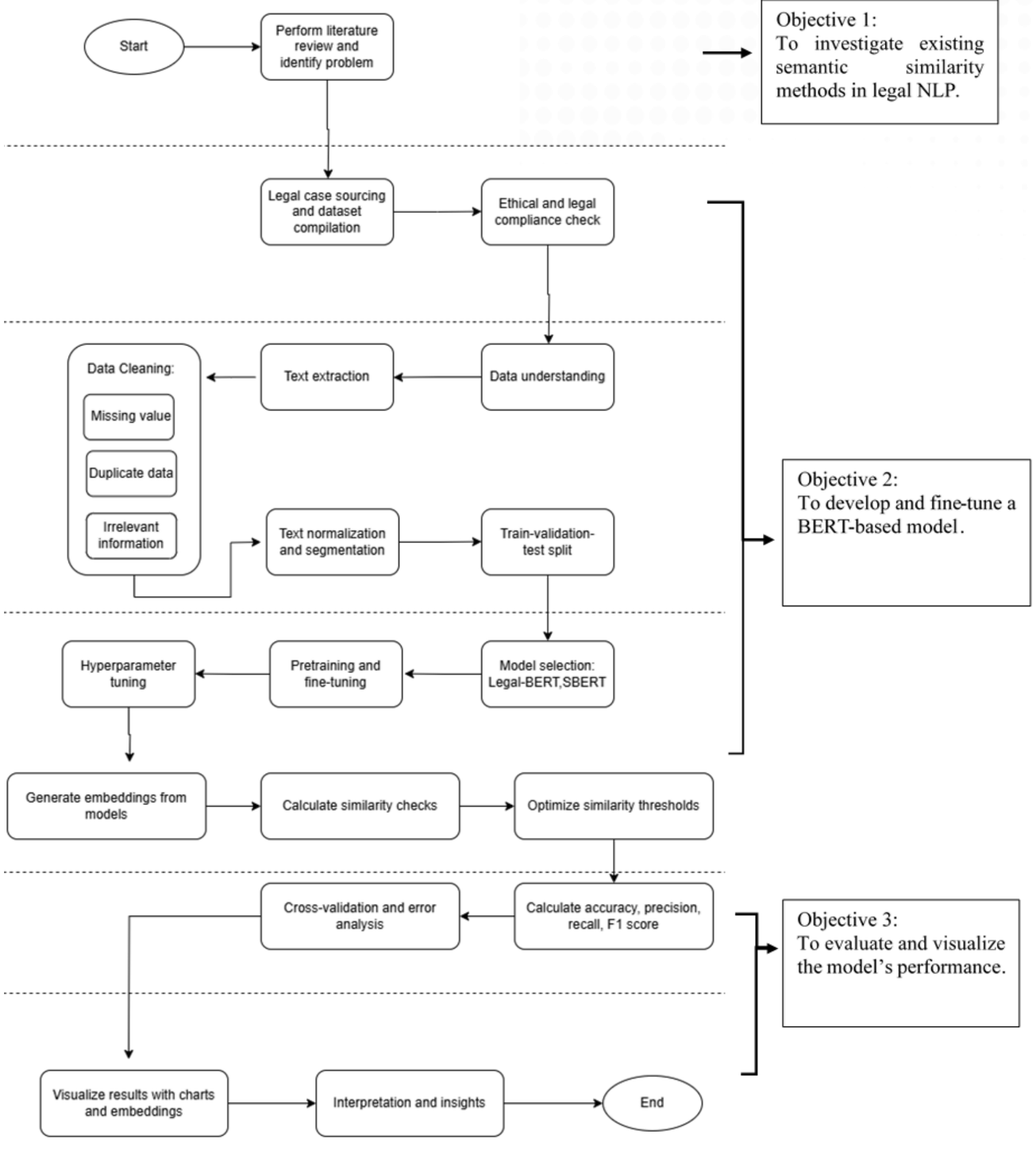
Data Acquisition

Data Preprocessing

Model Development

Evaluation & Validation

Visualization & Interpretation



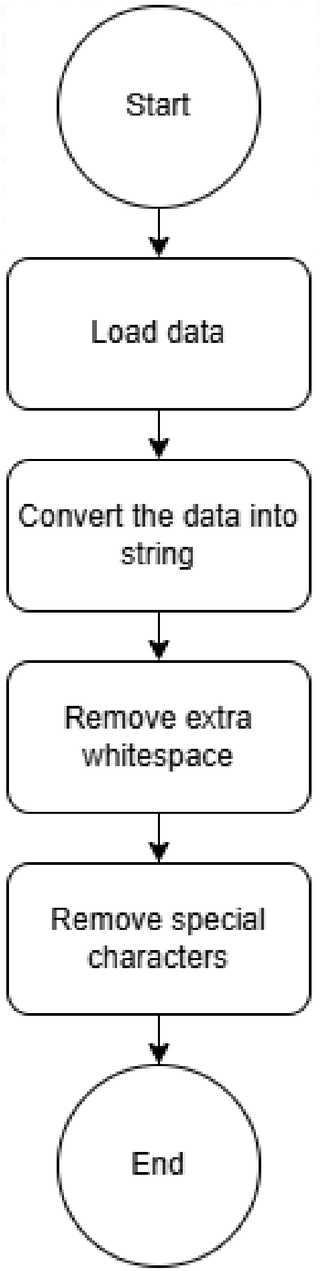
RESEARCH METHODOLOGIES

Dataset Overview

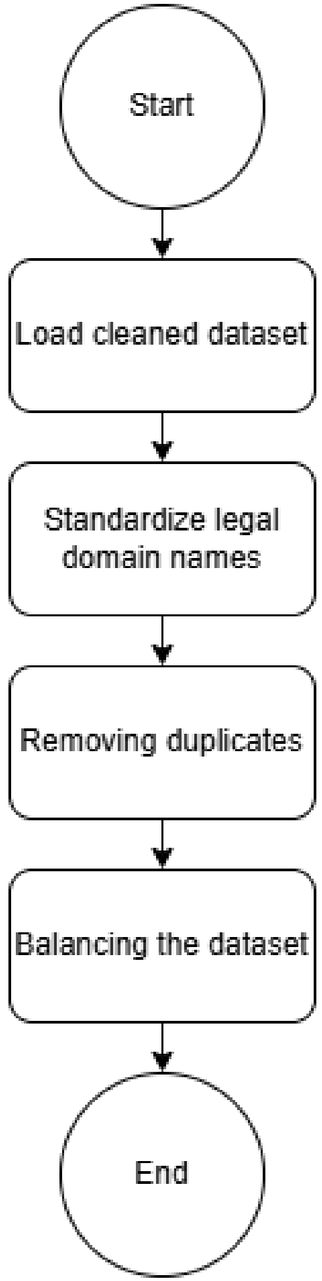
- **Data Source:** The dataset was curated from legal cases sourced from Kaggle.
- **Corpus Size:** Comprises of 3000 rows of a diverse set of legal cases, forming the basis for the model training and evaluation.

	A	B	C	D	E	F	G	H	I	J	K
1	id	case1_id	case2_id	case1_text	case2_text	case1_domain	case2_domain	true_label	is_ambigu	complexity	
2	1	ML-2024-2	KL-2019-4	Dispute arose ove	Partnership dissol	contract	commercial	0	FALSE	expert	
3	2	KL-2019-8	PG-2020-7	Dispute over owne	Strata property ca	property	property	1	FALSE	expert	
4	3	ML-2022-1	KL-2019-8	Adoption petition	Divorce proceedin	family	family	1	FALSE	expert	
5	4	PG-2021-2	ML-2020-5	Criminal intimidati	Criminal breach of	criminal	criminal	1	FALSE	expert	
6	5	PG-2023-9	PG-2018-3	Construction cont	Religious law conf	contract	constitutional	0	FALSE	expert	

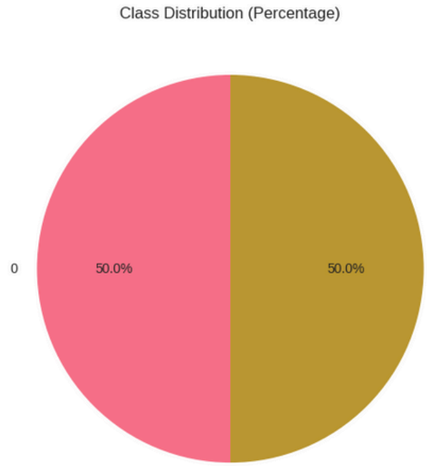
Dataset Cleaning Flowchart



Dataset Preprocessing Flowchart

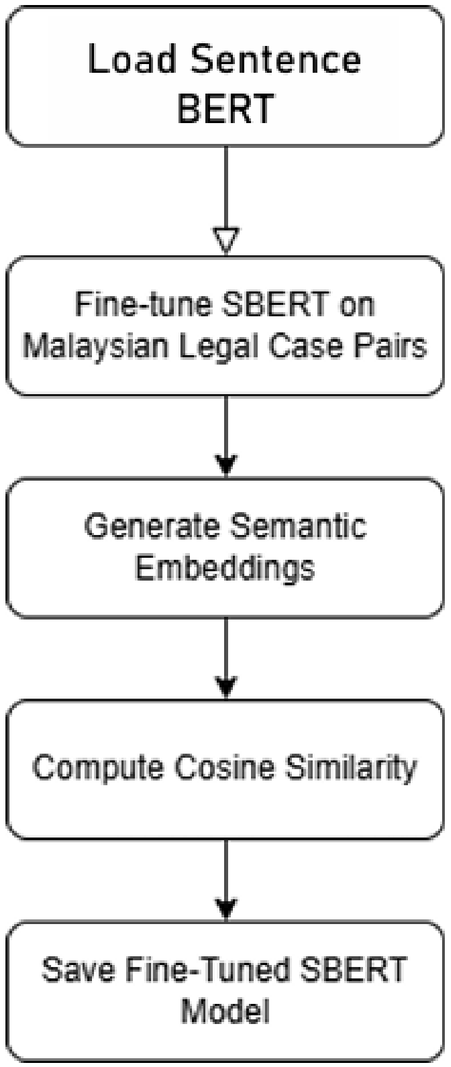


Descriptive statistics of data after cleaning and processing




```
--- Descriptive Statistics for Random Oversampled Data (Numerical Columns) ---
                id   true_label  case1_text_length  case2_text_length
count  2996.000000  2996.000000        2996.000000        2996.000000
mean   1481.342123     0.500000         286.278371         286.670895
std     866.066708     0.500083          29.171108          30.019345
min       1.000000     0.000000         237.000000         237.000000
25%     727.750000     0.000000         266.000000         265.000000
50%    1472.500000     0.500000         281.000000         281.000000
75%    2233.250000     1.000000         300.000000         299.000000
max    3000.000000     1.000000         418.000000         430.000000
```

Model Development



- **Load SBERT:** Start with a pre-trained SBERT model.
- **Fine-tune on Legal Cases:** Adapt SBERT using cleaned and processed Malaysian legal case pairs data to understand legal language.
- **Generate Embeddings:** Convert legal texts into semantic embeddings (numerical representations of meaning).
- **Compute Cosine Similarity:** Measure the similarity between legal texts by comparing their embeddings.
- **Save Model:** Preserve the fine-tuned SBERT for future use in identifying similar legal cases.

Experimental Setup

To determine the most effective approach, a comparative experiment was designed:

Model Category	Model Name	Description
Baseline (Traditional)	TF-IDF with Cosine Similarity	A statistical method based on word frequency.
Baseline (Traditional)	Bag-of-Words (BoW)	A simple word count representation.
Transformer-Based	Sentence-BERT (SBERT)	Fine-tuned all-MiniLM-L6-v2 model.

- **Similarity Calculation: Cosine Similarity was used to measure the distance between sentence embeddings for all models.**
- **Evaluation Task: The models were evaluated on their ability to classify pairs of legal text as "similar" or "not similar."**

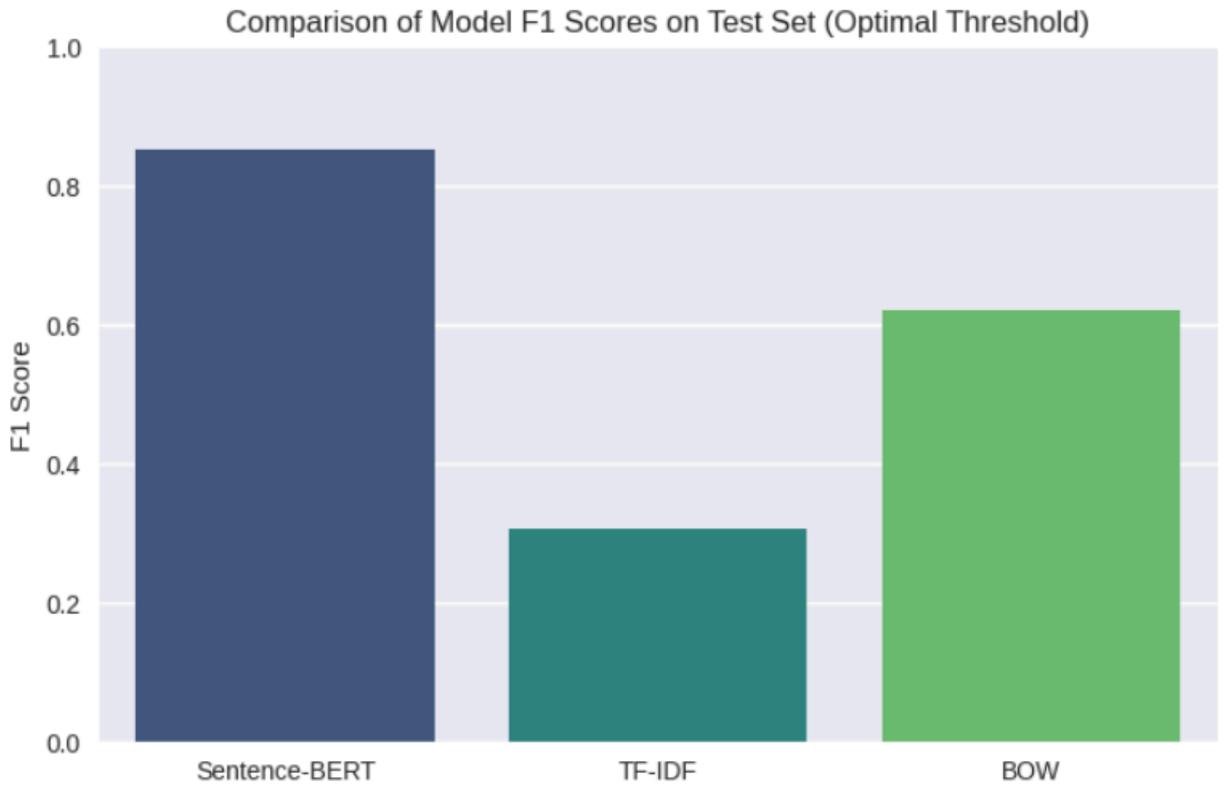
Results: Performance Metrics

The models were evaluated on Accuracy, F1-Score, Precision, and Recall.

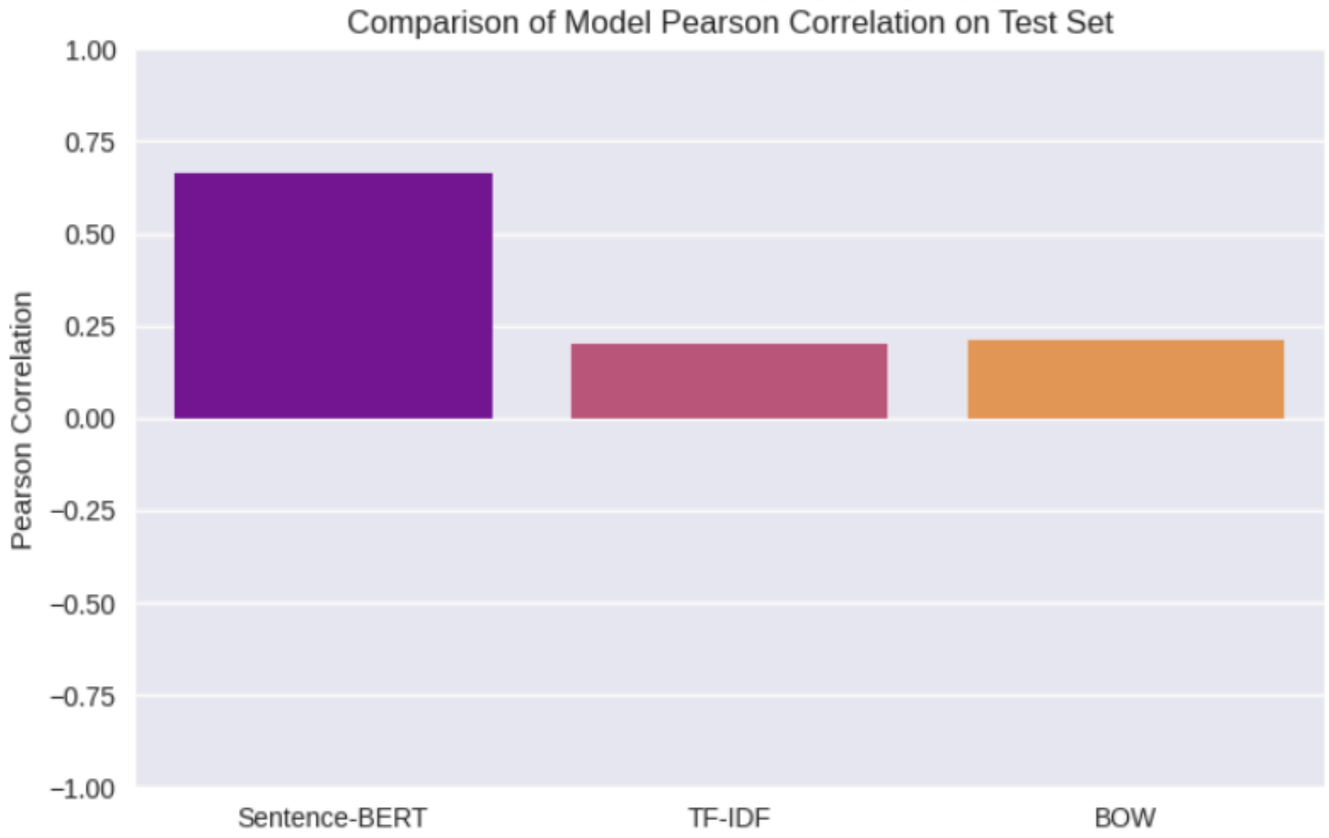
Model	Accuracy	F1-Score	Precision (Similar)	Recall (Similar)
TF-IDF	56.70%	0.3085	0.7632	0.1933
BoW	58.00%	0.6205	0.5659	0.6867
SBERT (all-MiniLM-L6-v2)	**83.00%**	**0.8523**	**0.75**	**1**

- The fine-tuned SBERT model drastically outperforms the traditional baseline models across all major metrics.
- TF-IDF showed extremely poor recall (0.19), failing to identify over 80% of relevant pairs.
- SBERT's perfect recall (1.0) indicates it successfully identified all of the truly similar pairs in the test set, making it highly reliable for discovery.

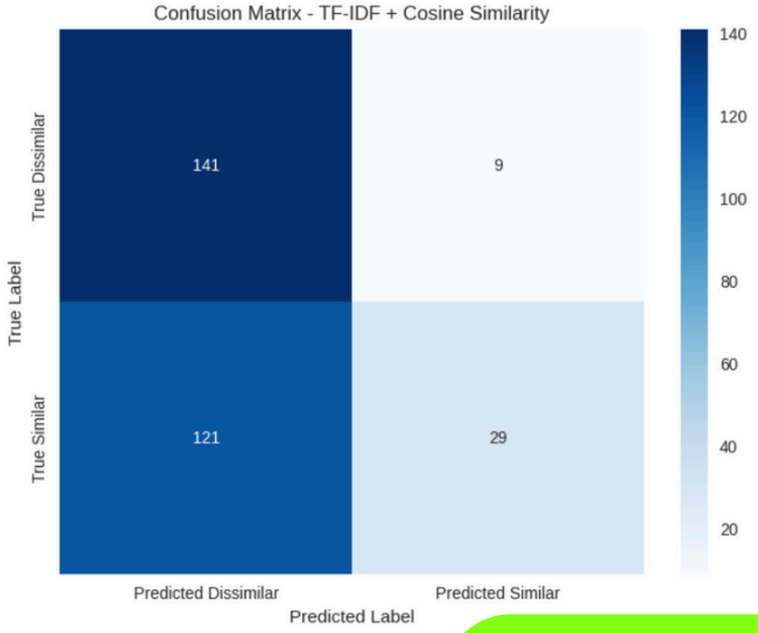
Quantitative Analysis: Model Comparison

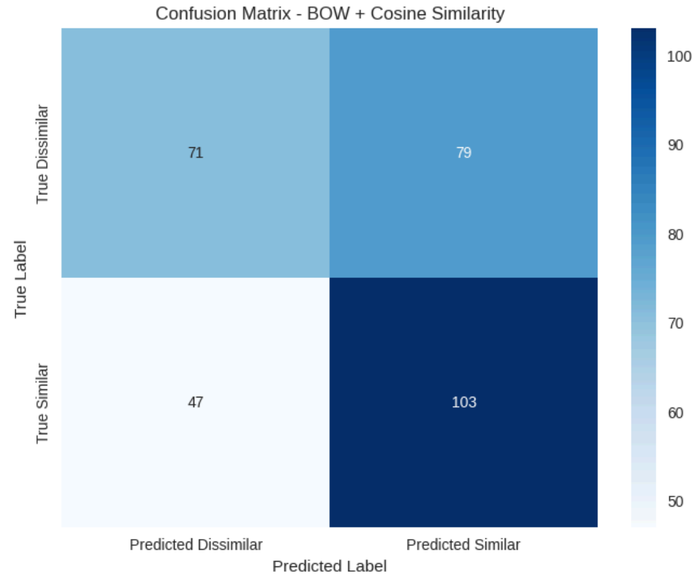


Quantitative Analysis: Model Comparison

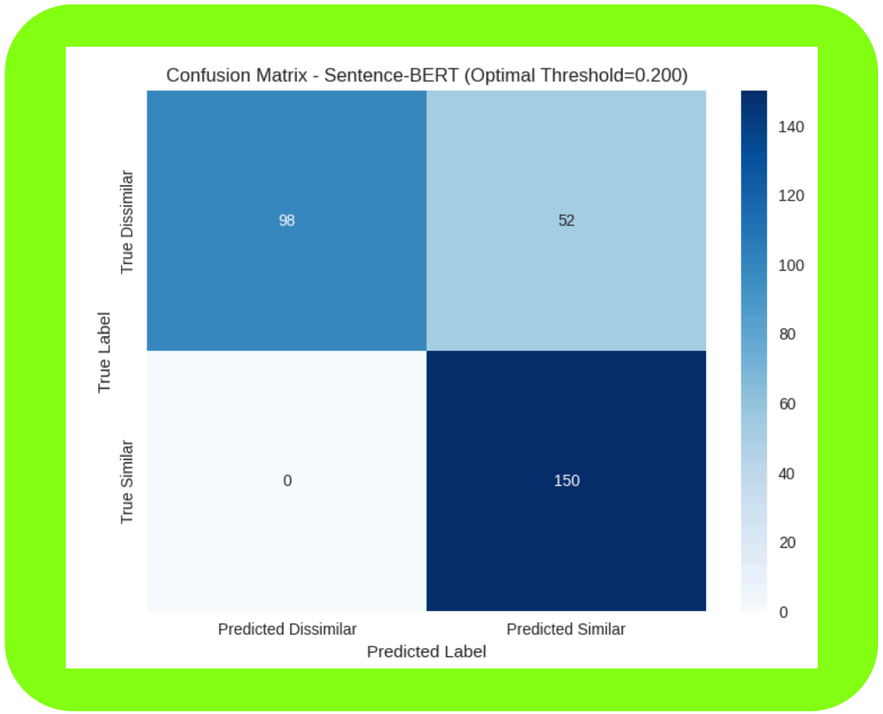


Result: Visual Analysis

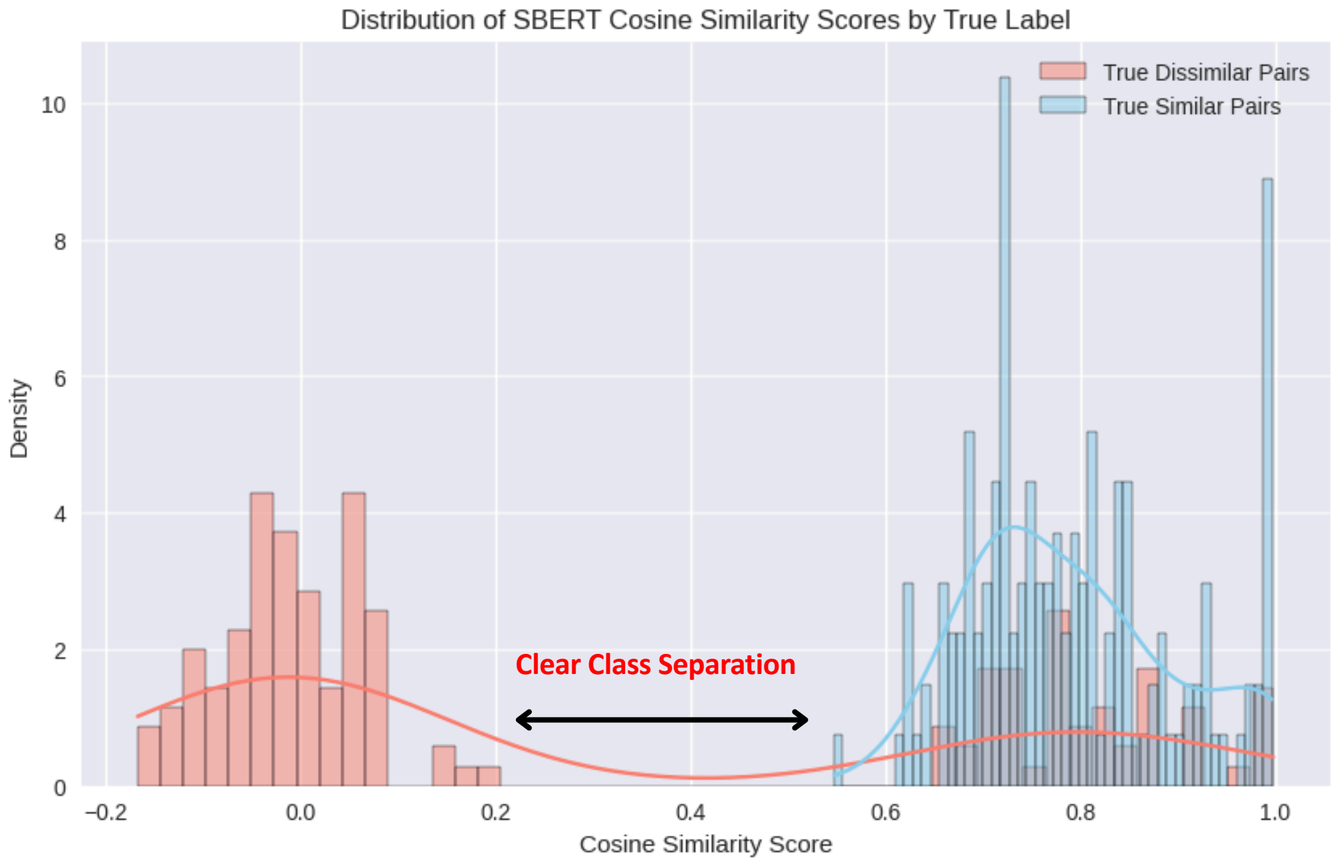




BEST RESULT



Result: Visual Analysis



Discussion of Results

Superiority of Contextual Embeddings:

The performance gap confirms that understanding context is crucial. The SBERT model's F1-score of 0.8523 starkly contrasts with 0.6205 for BoW and a mere 0.3085 for TF-IDF, proving that contextual models learn the nuances of legal language far more effectively.

High Recall is Critical for Legal Practice:

The model's perfect recall (1.0) is its most important feature. In legal research, failing to find a relevant precedent (a False Negative) can be catastrophic. The SBERT model acts as a powerful "safety net." In contrast, the TF-IDF model's recall of 0.1933 means it would miss four out of every five relevant precedents, a risk unacceptable in legal practice.

Practical Viability:

The combination of 83% accuracy and an F1-Score over 0.85 indicates that the model provides a reliable balance of precision and recall, making it a practically viable tool for real-world application.

Conclusion & Future Works

Achievements for this research are:

- This project successfully developed and validated a high-performance Sentence-BERT model fine-tuned for Malaysian legal text. It has been demonstrated to be a highly effective tool for semantic similarity, significantly overcoming the limitations of traditional search methods.

Future works in MDS Project II are:

- **Larger Domain-Specific Pre-training:** Further fine-tune the model on a much larger and more diverse corpus of Malaysian legal documents, including different court levels and languages.
- **Sentence-Pair Expansion:** Implement advanced techniques to generate more positive and hard-negative training pairs to further improve model nuance.
- **Contextual Embedding with Metadata:** Incorporate case metadata (such as judges, date, and cited statutes) as additional features to enrich the embeddings and improve relevance ranking.
- **Multilingual and Cross-Lingual Capabilities:** Expand the model to handle cases written in Bahasa Melayu and to find relevant precedents across languages.

THANK YOU

univteknologimalaysia utm.my utmofficial