Multi-class Medical Severity Classification (TF-IDF, GloVe, BERT)

ITCS348 INTRODUCTION TO NATURAL LANGUAGE PROCESSING

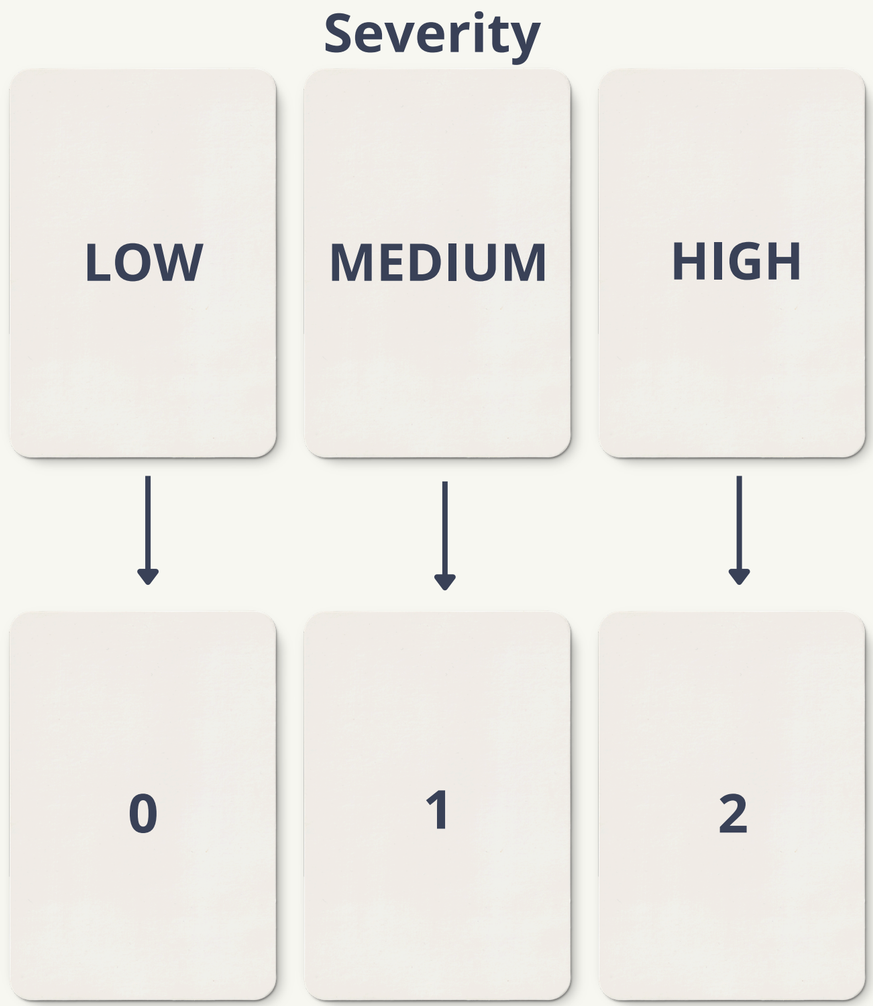
GROUP MEMBERS:
6688012 MR. PERAWIS BURANASING
6688121 MISS. SUPITSARA TANASARNSUKSTID
6688193 MR. ATICHAT KANGSAMUT

Motivation & Task (What we solve)

- Online health platforms need **early triage** from short patient text 📝
- Goal: classify severity into **Low / Medium / High** 🎯
- This is **decision-support**, not a replacement for clinicians 🚑

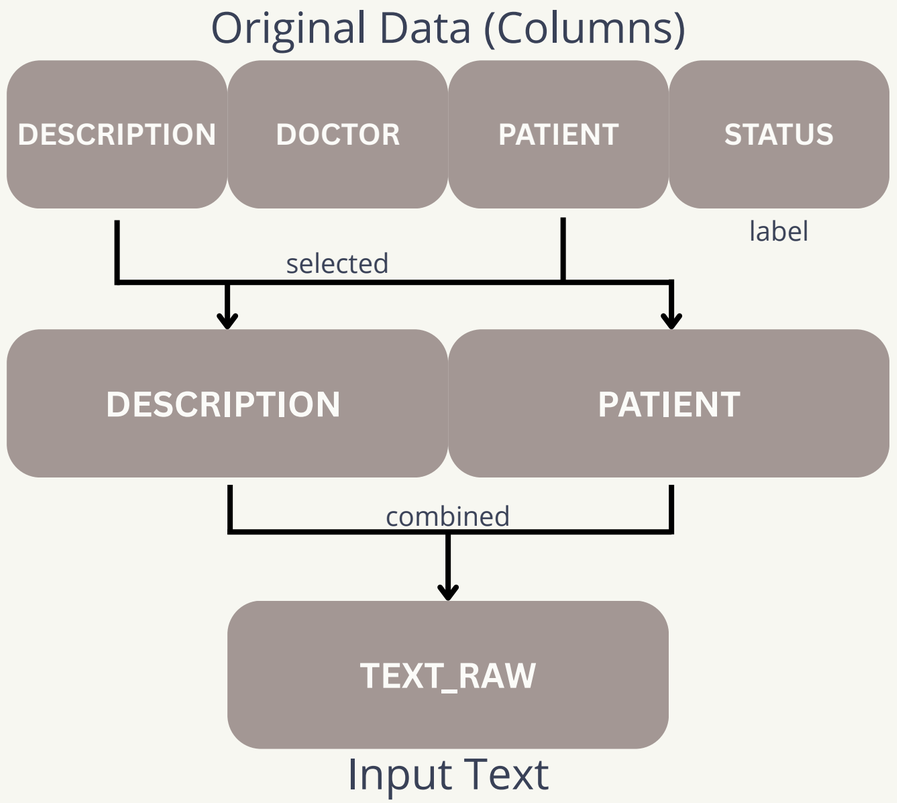
Labels + Why Macro Metrics

- 3-class classification: **low→0, medium→1, high→2**
- Dataset is **imbalanced**, so we report: Accuracy + **Macro-Precision / Macro-Recall / Macro-F1**
- Macro metrics treat each class equally → important for **high severity** detection



Dataset & Input Construction

- Source: Patient–Doctor Conversation Dataset
- Loaded from: hf://datasets/mahfoos/Patient-Doctor-Conversation/pred_status.csv
- Columns: Description, Doctor, Patient, Status
- Input text: text_raw = Description + Patient
- Total used (3-class): 3320 samples



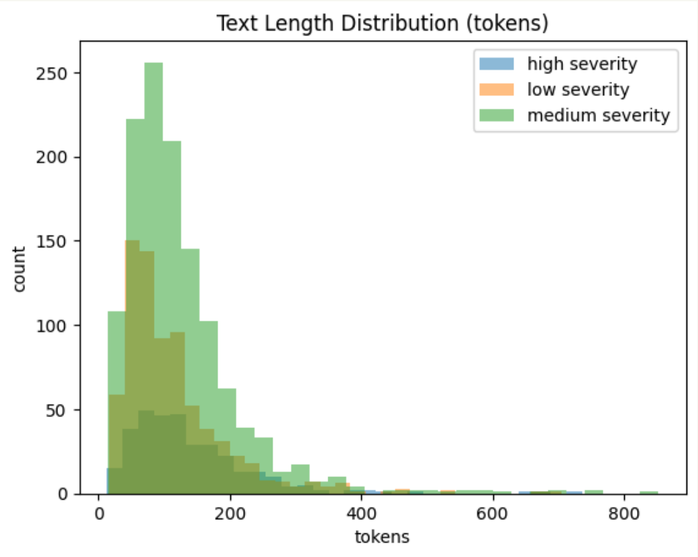
Split + Label Distribution + Length (EDA #1)

- Stratified split (70/15/15): Train 2323 / Val 498 / Test 499
- Class proportions: medium 53.58%, low 32.17%, high 14.25%
- Text length varies a lot → truncation matters for transformers

Label Distribution

Class	Total (n, %)	Train (n, %)	Val (n, %)	Test (n, %)
low severity	1068 (32.17%)	747 (32.16%)	160 (32.13%)	161 (32.26%)
medium severity	1779 (53.58%)	1245 (53.59%)	267 (53.61%)	267 (53.51%)
high severity	473 (14.25%)	331 (14.25%)	71 (14.26%)	71 (14.23%)

Text length distribution



EDA #2: Vocabulary + PMI-like Keywords

Vocabulary stats (vocab_size, TTR)

Class	num_docs	vocab_size	TTR
high severity	331	4852	0.105667
low severity	747	6143	0.072722
medium severity	1245	8385	0.054245

Top 3 PMI-like keywords per class (Example)

class	words
high severity	“pain”, “blood”, “test”, “months”, “back” (among top words)
low severity	“pain”, “last”, “time”, “day”, “back” (among top words)
medium severity	“pain”, “days”, “normal”, “back”, “last” (among top words)

- Vocabulary differs by class (medium has largest vocab; high has highest TTR)
- Many frequent words overlap → context matters
- PMI-like keywords show class-specific terms

Preprocessing + Ablation (Text Cleaning)

- We tested 4 preprocessing variants:
  A_raw, B_clean, C_stop, D_stop_lemma

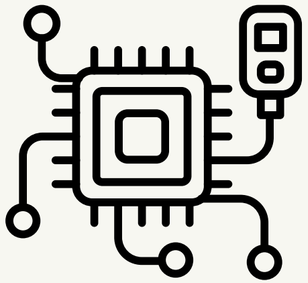
- Best validation macro-F1:
  A_raw and B_clean (tie)

- Decision: keep preprocessing minimal for main model comparisons

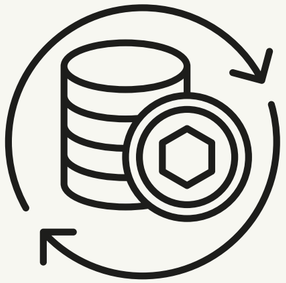
Ablation macro-F1 comparison plot

Experiment	macro-F1	vocab_size	vocab_reduction_%
A_raw	0.47909	31660	0
B_clean	0.47909	31660	0
D_stop_lemma	0.468989	18699	40.9381
C_stop	0.462282	18137	42.7132

Feature Extraction (TF-IDF vs GloVe vs BERT)



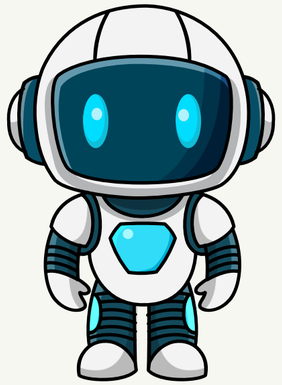
- TF-IDF n-grams: sparse vectors for LR/SVM/NB
- GloVe mean (100d): dense vectors for FNN/MLP
- BERT tokens: contextual input, **max length 512**



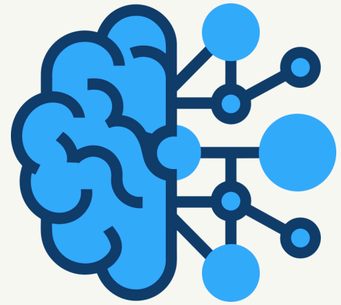
Feature representation summary

Representation	Used by	Dimensionality	Sparsity	Notes on memory/size
TF-IDF(1–2), min_df=2	LR/SVM/NB (baseline)	31660	~99.56%	Sparse, moderate vocab
TF-IDF(1–3), min_df=2	feature comparison	50113	~99.66%	Sparse, larger vocab
TF-IDF(1–3), min_df=1	LR tuned	318081	~99.91%	Very large sparse vocab
TF-IDF(1–1), min_df=2	NB tuned	6079	~98.84%	Smaller vocab than LR tuned
GloVe mean (100d)	FNN/MLP	100	0% (dense)	Compact dense vectors
BERT tokens	BERT	≤512 tokens	n/a	Contextual, truncation at 512

Models Implemented (3+ models requirement)



- ML (TF-IDF): Logistic Regression, Linear SVM, Multinomial NB
- DL baseline: FNN/MLP (GloVe mean, 100d)
- Transformer: BERT fine-tuning (bert-base-uncased)



Classical ML → **Deep Learning** → **Transformer**

Validation Results + Tuning (structured experiments)

- Validation: BERT best macro-F1 0.5055
- Strongest classical baseline: LR macro-F1 (0.4791)
- Tuning: GridSearchCV (LR/SVM/NB) + BERT LR/batch size tuning

Validation results (all models)

Model	Accuracy	Macro-Precision	Macro-Recall	Macro-F1
Logistic Regression	0.540161	0.4772	0.481675	0.47909
Linear SVM	0.534137	0.424307	0.408701	0.405891
Multinomial NB	0.536145	0.178715	0.333333	0.23268
FNN (GloVe mean)	0.564257	0.385527	0.379198	0.337035
BERT	0.578313	0.528295	0.494542	0.505512

BERT tuning grid + best config

lr	batch_size	epochs	val_accuracy	val_macro_f1	val_loss
0.00003	16	10	0.580321	0.518154	1.484953
0.00002	8	10	0.616466	0.506041	0.880834
0.00002	16	10	0.590361	0.498468	1.075962
0.00003	8	10	0.604418	0.472621	0.894638

Tuned ML results + best params

Model	Best Params	Accuracy	Macro-Precision	Macro-Recall	Macro-F1
LR (tuned)	C=0.1, ngram=(1,3), min_df=1	0.522088	0.477261	0.487885	0.476687
Linear SVM (tuned)	C=0.1, ngram=(1,2), min_df=2	0.572289	0.497061	0.43514	0.436589
Multinomial NB (tuned)	alpha=0.1, ngram=(1,1), min_df=2	0.600402	0.575954	0.431212	0.41614

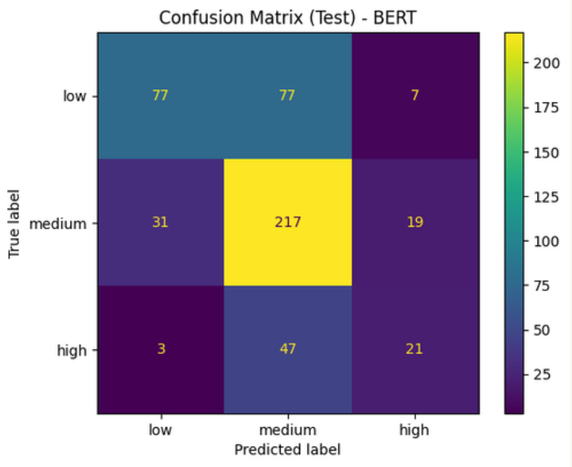
Test Results + Confusion Matrix + Per-class F1

- Test: BERT is best (Accuracy 0.6313, Macro-F1 0.5453)
- Key error: High → Medium (under-triage risk)
- Per-class: “High” is hardest class

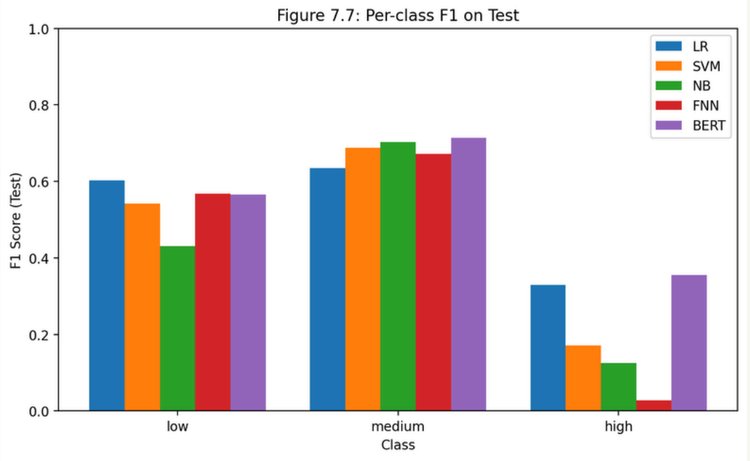
Test results (all models)

Model	Accuracy	Macro-Precision	Macro-Recall	Macro-F1
LR	0.571142	0.521302	0.533836	0.52261
SVM	0.593186	0.506236	0.462202	0.467086
NB	0.589178	0.587267	0.427837	0.419425
FNN	0.589178	0.717408	0.452683	0.422811
BERT	0.631263	0.592289	0.528923	0.545308

Confusion matrix (BERT test)



Per-class F1 on Test



Demo (Qualitative) + Cost + Interpretability + Ethics + Conclusion

Qualitative demo: show misclassified examples (why hard)

Misclassified test examples

true_name	pred_name	conf	snippet	explanation
low	medium	0.999423	on taking medications my postnatal drip has become thick why hello doctor I am having postnatal…	Class overlap
high	medium	0.999421	I have headache and increased heart rate with increased get and alt what is my problem hi doctor I…	Class overlap

Cost trade-off: BERT best but expensive

Cost comparison

model_key	train_time_sec	num_params	model_size_mb	inf_time_sec	samples_per_sec
LR	4.462953	1114437	19.11052	0.127847	3903.101
SVM	1.009071	110970	1.68222	0.073881	6754.062
NB	0.327901	20058	0.440515	0.034484	14470.52
FNN	1.52928	26627	0.101574	0.000464	1076339
BERT	490.7962	109000000	417.6504	8.30108	60.11266

Interpretability: NB / LR / SVM show indicative tokens/weights

NB top indicative tokens

Class	Token	Score	Interpretation
Low	retainer	1.8323	Dental/orthodontic term, usually routine care.
Medium	shivering	1.2617	Symptom that may indicate infection or systemic issue but not always critical.
High	suicidal	2.2262	Strong indicator of severe psychiatric emergency (high-risk case).

LR top weighted TF-IDF features

Model	Class	Top Positive Feature	Weight
LR	Low	skin	0.2121
LR	Medium	pain	0.2302
LR	High	cancer	0.2777

SVM top weighted TF-IDF features

Model	Class	Top Positive Feature	Weight
SVM	Low	teeth	0.5744
SVM	Medium	pain	0.5619
SVM	High	cancer	0.9393

Ethics

Under-triage risk → human-in-loop + abstain option

Conclusion

Main difficulty is adjacent-class separation (medium vs high)

Thank you!