

Crisis Severity Detection with TF-IDF Logistic Regression and DistilBERT: A Cost-Sensitive Approach with Lightweight Defense-Oriented Adaptation

Donghyun Kim
Department of Statistics
University of Michigan
Ann Arbor, MI, USA
donghki@umich.edu

Abstract—This project uses the CrisisMMD corpus as a proxy to construct a four-level crisis severity detection system and evaluates three modeling approaches: a TF-IDF + Logistic Regression baseline, a fine-tuned DistilBERT classifier, and a lightweight domain-adapted DistilBERT variant based on a custom defense-oriented lexicon. DistilBERT substantially outperforms the classical baseline and reduces cost-sensitive risk. Lexicon-weighted domain adaptation alters model behavior but does not improve overall performance, highlighting both the strength of Transformer models and the limitations of simple oversampling for domain specialization.

Index Terms—text classification, transformers, domain adaptation, triage

I. INTRODUCTION

Crisis-response and defense logistics organizations routinely process large volumes of short, noisy textual reports describing casualties, infrastructure damage, displaced populations, and disruptions to transportation or supply networks. Although these messages may appear similar, they differ substantially in the type of response required. Many incidents fall under the responsibility of civilian authorities, whereas others demand rapid military involvement. Even within military contexts, procedures for force posture, sustainment, and escalation vary depending on the assessed level of urgency. Misjudging either the severity or the operational relevance of an event can delay the appropriate response pathway, creating risks for both civilian populations and military units.

Authentic operational military reports are not publicly available due to security restrictions, so research on military-relevant text triage must rely on public proxy datasets that contain crisis-related content. Transformer architectures such as DistilBERT have become state-of-the-art for text classification [1], although their use in operational triage or defense logistics remains underexplored. Meanwhile, triage systems in clinical and emergency settings (e.g., the Emergency Severity Index), illustrate the value of ordered severity scales and asymmetric penalties for underestimation [2]. These concepts, however, have not been widely integrated into crisis-text modeling.

This creates two modeling challenges: (i) constructing a severity scale aligned with operational decision-making, and (ii) adapting general-purpose language models to be more attentive to defense-related linguistic cues without access to labeled in-domain data. The goal of this project is to address these challenges by defining an operationally inspired four-level severity mapping for crisis messages and evaluating three modeling approaches on the CrisisMMD dataset: a TF-IDF + Logistic Regression baseline, a fine-tuned DistilBERT model, and a lightweight domain-adapted variant of DistilBERT using lexicon-weighted oversampling. Concretely, the study investigates two questions: (i) whether a Transformer-based model meaningfully shows better performance than a classical keyword-driven baseline for crisis severity classification, especially under class imbalance and for rare high-severity events; and (ii) how lexicon-guided domain adaptation, applied to the same Transformer architecture, changes both the performance and behavior of the model on defense-relevant crisis text.

II. METHOD

A. Exploratory Data Analysis

The humanitarian text subset of Crisis MMD consists of short English tweets annotated into eight humanitarian information types. Initial inspection revealed substantial class imbalance: labels such as *not_humanitarian* and *other_relevant_information* dominate the dataset, whereas labels such as *missing_or_found_people* and *vehicle_damage* are rare. Most tweets contain fewer than twenty words and exhibit noise, abbreviations, and high lexical variability. These observations are illustrated in Figs. 1 and 2.

B. Severity Mapping: Design and Rationale

Because the original CrisisMMD labels represent information types rather than urgency levels, they are not suitable for triage-style or operational decision-making. To construct an ordered severity target, the eight humanitarian categories are mapped into four operational levels: Routine, Elevated,

Urgent, and Critical. Routine includes background or non-humanitarian content; Elevated includes infrastructure or utility damage, vehicle damage, and rescue or donation-related efforts; Urgent corresponds to reports involving affected individuals; and Critical includes cases describing injuries, fatalities, or missing persons. This mapping is conceptually inspired by multi-level urgency structures found in military medical evacuation precedence (e.g., U.S. Army medical doctrine [3]) and clinical triage systems [2].

C. TF-IDF + Logistic Regression Baseline

The classical baseline model uses TF-IDF vectorization to convert tweets into sparse numerical features. Preprocessing includes lowercasing and tokenization, followed by unigram and bigram extraction with a vocabulary cap of 10,000 terms and a minimum document frequency of two. A multinomial Logistic Regression classifier is trained with the *lbfgs* optimizer, using inverse-frequency class weights to mitigate imbalance. This model serves as a keyword-driven reference for evaluating improvements from contextual representations.

D. DistilBERT Baseline

The primary contextual model is DistilBERT, fine-tuned as a four-class sequence classifier. Tweets are tokenized using the DistilBERT tokenizer with a maximum sequence length of 64 tokens. Fine-tuning is performed using the Hugging Face Transformers library with standard hyperparameters, including weight decay of 0.01, batch sizes of 16 for training and 32 for evaluation, and three training epochs.

E. Domain-Adapted DistilBERT (Lexicon-Weighted Variant)

To investigate lightweight domain adaptation, a defense-oriented lexicon was constructed from publicly available military logistics doctrine [4]–[6] and humanitarian logistics literature [7]. The resulting lexicon contains several dozen terms related to logistics infrastructure (e.g., roads, bridges, ports), sustainment resources (e.g., fuel, supplies, depots), vehicles and equipment (e.g., trucks, armored vehicles), and unit-level operational contexts (e.g., checkpoints, battalions).

Tweets containing at least one lexicon term are tagged as defense-like and oversampled during training by duplicating each such tweet three additional times, while validation and test sets remain unchanged. A fresh DistilBERT model is then fine-tuned on this augmented training set using the same hyperparameters as the baseline. This approach preserves the model architecture and label space while altering only the sampling distribution, providing a lightweight form of domain adaptation that increases the model’s exposure to defense-relevant language.

F. Evaluation Protocol

All models are evaluated on identical validation and test splits for direct comparability. Standard metrics include accuracy, macro-F1, per-class precision/recall, and confusion matrices. Because crisis and defense operations involve asymmetric risks—particularly when high-severity events are

underestimated—a custom cost-sensitive evaluation is also employed. A 4×4 cost matrix C is defined with zero cost for correct predictions, high penalties for underestimating Urgent or Critical events (e.g., predicting Routine when the true label is Critical), and lower but non-zero penalties for false alarms and mid-range confusions. The specific matrix used in this project is:

$$C = \begin{bmatrix} 0 & 1 & 3 & 4 \\ 1 & 0 & 2 & 4 \\ 3 & 2 & 0 & 2.5 \\ 5 & 4 & 2.5 & 0 \end{bmatrix},$$

where rows correspond to the true labels (Routine, Elevated, Urgent, Critical) and columns correspond to predicted labels in the same order. The structure reflects an operational preference for avoiding missed high-severity incidents: the highest penalties occur when Critical or Urgent events are misclassified as low-severity (e.g., Critical \rightarrow Routine), while false alarms incur smaller penalties.

Given confusion matrix counts n_{ij} , total risk is computed as

$$R_{\text{total}} = \sum_{i,j} n_{ij} C_{ij},$$

with average risk defined as $R_{\text{avg}} = R_{\text{total}}/N$. This metric summarizes the expected penalty per prediction under the chosen operational risk model. Although it is designed to penalize severe underestimation more heavily, in practice it provides a complementary view to macro-F1 by capturing how each model distributes errors across severity levels.

III. RESULTS

After applying the four-level severity mapping to the CrisisMMD humanitarian labels, the new severity variable was merged back into the original train, validation, and test splits. Logistic Regression, the DistilBERT baseline, and the domain-adapted DistilBERT all used these mapped splits to ensure fair comparison, with each model trained only on the training set and evaluated on the fixed test split. The results are summarized in Fig. 3 and further detailed in Figs. 4–9.

The first comparison, Logistic Regression versus DistilBERT, isolates the effect of model architecture. The DistilBERT baseline outperformed the Logistic Regression model across all primary metrics: test accuracy increased from 0.65 to 0.70, macro-F1 from 0.47 to 0.52, and average cost-sensitive risk decreased from 0.58 to 0.42. Per-class scores indicate that these gains are driven mainly by more stable predictions for the majority classes (Routine and Elevated) and modest improvements for the sparse Critical class, while recall for Urgent events remains low in both models. The reduction in average risk is influenced partly by the slight improvement in Critical recall, but primarily by fewer low- and mid-severity misclassifications, whose large sample counts contribute substantially to the overall risk despite their lower penalties. Severe underestimation of Urgent and Critical events therefore remains a key limitation across both models.

The second comparison examines whether lightweight domain adaptation can further improve the DistilBERT baseline. The domain-adapted model achieved slightly lower overall performance, with test accuracy of 0.6853, macro-F1 of 0.4760, and average risk of 0.4613. Confusion matrices show that many Routine samples were reassigned to higher predicted severity levels, particularly Elevated, while recall for Critical events dropped from about 0.67 to 0.41. This indicates that lexicon-guided oversampling successfully pushed the model to react more strongly to defense-related terms, biasing its decision boundaries upward in predicted severity.

However, this shift did not generalize to the full CrisisMMD distribution. The increased tendency to up-classify Routine samples created additional mid-severity confusions, and the decline in Critical recall means that high-impact errors became more frequent. Under the cost matrix used in this project—which assigns the highest penalties to underestimating Urgent and Critical cases—the net effect is a slight increase in average risk. Thus, while the domain-adapted model behaves differently, the oversampling strategy did not yield improvements aligned with the operational priorities encoded in the cost-sensitive evaluation.

IV. CONCLUSION

This project evaluated four-level severity classification using Logistic Regression, DistilBERT, and a lexicon-guided domain-adapted variant. Although DistilBERT reduced average risk and improved overall metrics, both DistilBERT and Logistic Regression models continued to struggle with rare high-severity events, indicating that current approaches are insufficient for reliably identifying Critical and Urgent cases. The cost-sensitive metric also showed that numerical improvements in average risk do not consistently reflect reductions in the most consequential underestimation errors, underscoring the need for refined cost-matrix design and targeted handling of sparse classes.

The lexicon-based domain adaptation method shifted model behavior—often raising predicted severity for defense-related language—but did not improve overall performance and, in some cases, increased mid-level confusion. More effective adaptation will likely require stronger techniques such as sample weighting, continual pretraining, or multi-task learning. Despite these limitations, the study shows that public crisis datasets can serve as useful proxies for triage-style modeling and highlights both the potential and challenges of adapting Transformer models for defense-relevant decision-support tasks.

REFERENCES

- [1] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter,” arXiv:1910.01108, 2019.
- [2] N. Gilboy, T. Tanabe, D. Travers, and R. Wuerz, *Emergency Severity Index (ESI): A Triage Tool for Emergency Department Care, Version 4, Implementation Handbook*. Rockville, MD, USA: Agency for Healthcare Research and Quality, 2005.
- [3] U.S. Department of the Army, *ATP 4-02.2: Medical Evacuation*. Washington, DC, USA: Headquarters, Department of the Army, 2019.

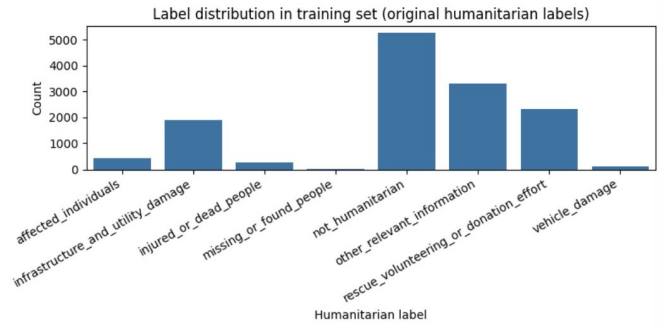


Fig. 1. Label distribution in CrisisMMD

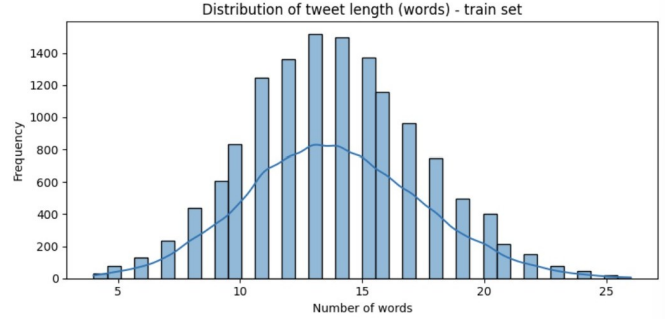


Fig. 2. Distribution of tweet lengths in CrisisMMD

- [4] NATO Standardization Office, *AJP-4: Allied Joint Doctrine for Logistics*. NATO Standardization Office, 2018.
- [5] Joint Chiefs of Staff, *JP 4-0: Joint Logistics*. U.S. Department of Defense, 2019.
- [6] NATO Standardization Office, *AJP-4.4: Allied Joint Doctrine for Movement and Transportation*. NATO Standardization Office, 2013.
- [7] P. Tatham and L. Houghton, “The wicked problem of humanitarian logistics and disaster relief aid,” *Journal of Humanitarian Logistics and Supply Chain Management*, vol. 1, no. 1, pp. 15–31, 2011.
- [8] D. Kim, “Project_507,” GitHub repository, 2025. [Online]. Available: https://github.com/DonghyunKim0526/Project_507

	Model	Accuracy	Macro F1	Avg Risk	Total Risk
	TF-IDF + Logistic Reg.	0.6455	0.4722	0.5849	1308.5
	DistilBERT	0.7032	0.5221	0.4213	942.5
	Domain-Adapted DistilBERT	0.6853	0.4760	0.4613	1032.0

Fig. 3. Test Performance Summary

	precision	recall	f1-score	support
Routine	0.803	0.713	0.755	1427
Elevated	0.525	0.560	0.542	678
Urgent	0.099	0.186	0.130	86
Critical	0.357	0.652	0.462	46
accuracy			0.646	2237
macro avg	0.446	0.528	0.472	2237
weighted avg	0.682	0.646	0.661	2237

Fig. 4. Classification report for the Logistic Regression model

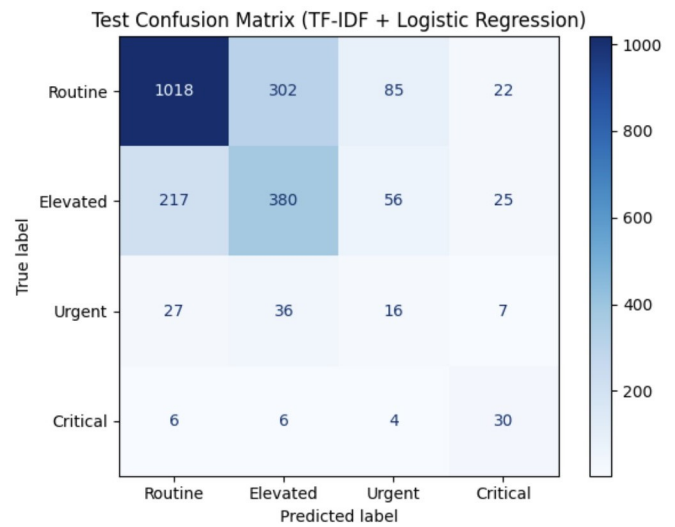


Fig. 7. Test confusion matrix for the Logistic Regression

Classification report:

	precision	recall	f1-score	support
Routine	0.785	0.805	0.795	1427
Elevated	0.565	0.566	0.566	678
Urgent	0.290	0.105	0.154	86
Critical	0.500	0.674	0.574	46
accuracy			0.703	2237
macro avg	0.535	0.538	0.522	2237
weighted avg	0.693	0.703	0.696	2237

Fig. 5. Classification report for DistilBERT

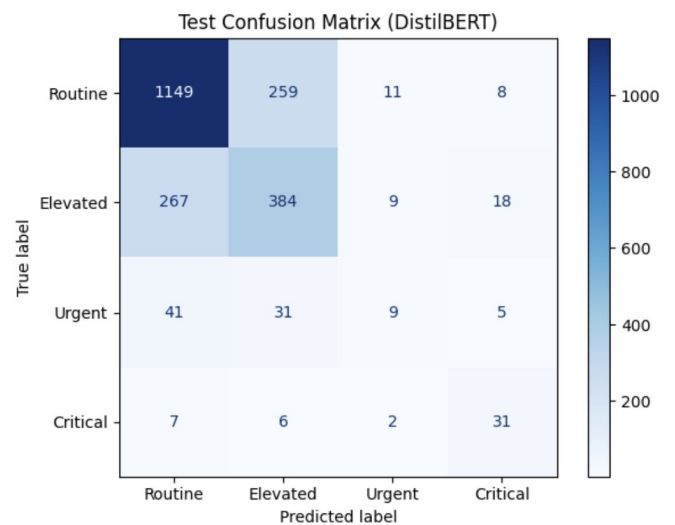


Fig. 8. Test confusion matrix for DistilBERT

Classification report:

	precision	recall	f1-score	support
Routine	0.780	0.783	0.781	1427
Elevated	0.541	0.571	0.556	678
Urgent	0.227	0.116	0.154	86
Critical	0.413	0.413	0.413	46
accuracy			0.685	2237
macro avg	0.490	0.471	0.476	2237
weighted avg	0.679	0.685	0.681	2237

Fig. 6. Classification report for Domain-Adapted DistilBERT

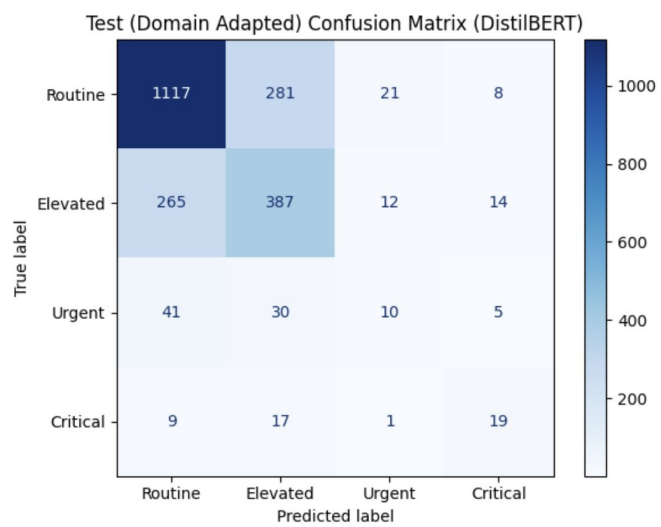


Fig. 9. Test confusion matrix for Domain-Adapted DistilBERT