**NLP for Tracking Misinformation in Low-Resourced Public Health Datasets**

**Agnivo Basu**Reg. No. 242808001  
MSc Digital Epidemiology

**Manu ML**Reg. No. 242807006  
MSc Data Science

**Subject: Deep Learning & Text Mining (DDS5205)**

***Department of Applied Statistics & Data Science,   
Prasanna School of Public Health,   
MAHE, Manipal***

**Date of Submission: 12th May, 2025**

**Abstract**

The rapid spread of health-related misinformation on digital platforms poses a critical threat to public well‐being, particularly in settings with limited annotated corpora and language‐specific NLP tools. This study aims to evaluate the effectiveness of transformer‐based models for detecting four categories of misinformation—Reliable, Fake, Unreliable, and “Not Enough Info”—in low‐resourced public health datasets. We first compiled and preprocessed a novel corpus of 8 927 English‐language instances by merging six publicly available health‐misinformation datasets, stratifying labels into a consistent four‐way taxonomy. Using Hugging-Face Transformers, we fine-tuned three architectures—DistilBERT, BERT, and RoBERTa—on an 80/20 train–test split, optimizing hyperparameters such as learning rate, batch size, and weight decay. Model performance was assessed via accuracy, precision, recall, F1-score, and cross-entropy loss, and embedding separability was visualized through 2D and 3D t-SNE projections and SHAP token-importance plots. DistilBERT achieved 75.7 % accuracy (macro-F1 = 0.74), BERT reached 82.1 % (macro-F1 = 0.82), and RoBERTa attained the highest accuracy of 85.7 % (macro-F1 = 0.80). While transformer embeddings distinctly clustered the “Not Enough Info” and “Unreliable” classes, “Fake” and “Reliable” remained partially overlapping, indicating semantic proximity. These findings demonstrate that deep contextual models substantially outperform lighter baselines but still face challenges in discriminating nuanced misinformation. Future work should explore data augmentation, ensembling, domain‐adaptive pre-training, and multilingual extensions. By enhancing automated detection of public health misinformation, our results offer a scalable foundation for real-time monitoring systems and targeted interventions in resource‐constrained environments.

**1. Introduction**

Health‐related misinformation—false or misleading information about disease prevention, diagnosis, or treatment—poses a serious threat to public well‐being and trust in medical authorities (Wardle & Derakhshan, 2017). The rapid proliferation of rumors and unverified claims on social media and messaging platforms accelerates the spread of such misinformation, particularly in low‐resource settings where annotated corpora and language‐specific tools are scarce. For example, a randomized survey on WhatsApp users in the UK and Brazil demonstrated that partial or wholly false COVID-19 messages strongly influenced younger adults’ beliefs and sharing intentions, even after corrective information was provided (Vijaykumar et al., 2021). Meanwhile, public health communications may unintentionally omit critical context—such as effect sizes or uncertainty—further complicating individuals’ ability to distinguish fact from fiction (Brown & de Barra, 2023).

Confronting this “infodemic” requires robust natural language processing (NLP) techniques tailored both to resource‐rich and low‐resourced languages and domains. Traditional machine‐learning classifiers (e.g., support vector machines (SVMs), logistic regression) have long leveraged handcrafted features and bag-of-words representations to detect false claims (Dai et al., 2020). More recently, word‐embedding and transformer models (e.g., Word2Vec, FastText, BERT) have enabled richer semantic encoding, offering enhanced discrimination of subtle linguistic cues in fake news (Al-Tarawneh et al., 2024; Oubenali et al., 2022). Yet these advanced architectures often demand extensive training data and computational resources, posing challenges for deployment in low-resource public health contexts. This study therefore investigates both traditional and deep‐learning approaches—fine-tuned on a multilingual public‐health corpus—to evaluate their relative effectiveness in tracking misinformation under constrained settings.

**2. Literature Review**

**2.1 Traditional Machine‐Learning Approaches**  
Early fake‐news detection relied on classical algorithms with sparse, manually engineered features. Dai et al. (2020) introduced FakeHealth, a repository combining article text with social‐engagement metadata and expert reviews, and showed that SVMs using term frequency–inverse document frequency (TF–IDF) features achieved strong baseline accuracy. Similarly, traditional classifiers such as naive Bayes, *k*‐nearest neighbors, and random forests have been widely applied to labeled health‐news corpora, often achieving acceptable performance on balanced datasets but struggling with nuanced or context‐dependent misinformation (Dai et al., 2020).

**2.2 Word‐Embedding Techniques**  
The advent of dense vector representations transformed feature extraction in text classification. Al-Tarawneh et al. (2024) systematically compared TF–IDF, Word2Vec, and FastText embeddings across several machine‐learning and deep‐learning models on the TruthSeeker dataset. They reported that SVMs paired with TF–IDF and convolutional neural networks (CNNs) fed by TF–IDF both outperformed their Word2Vec and FastText counterparts, underscoring that the choice of embedding critically affects downstream classifier efficacy. Oubenali et al. (2022) reviewed medical‐concept embeddings—Word2Vec, GloVe, FastText, and bidirectional encoder representations from transformers (BERT)—highlighting that embedding visualization is essential for evaluating semantic consistency and revealing clustering of clinical terms, yet noting a lack of standardized guidelines for such visual analyses.

**2.3 Deep‐Learning and Transformer Models**  
Transformer‐based architectures, pre–trained on large corpora and fine-tuned on task‐specific data, have set new state-of-the-art results in many NLP benchmarks. Models such as BERT, RoBERTa, and their domain‐adapted variants (e.g., COVID-Twitter-BERT) capture long‐range dependencies and contextual nuances that elude simpler models (Al-Tarawneh et al., 2024). However, these models’ computational demands and reliance on large annotated datasets present challenges in low-resource environments. Techniques such as knowledge distillation (e.g., DistilBERT) and parameter sharing offer lighter‐weight alternatives, though often at some loss in accuracy.

**2.4 Ethical and Communicative Considerations**  
Beyond algorithmic performance, the ethical framing and transparency of public health communication influence misinformation dynamics. Brown and de Barra (2023) developed a taxonomy of non-honesty in public health messages—ranging from omission of effect sizes to strategic framing—that underscores how even well-intentioned advisories can mislead. Vijaykumar et al. (2021) further demonstrated that corrective information must account for age‐related differences in source monitoring and cognitive processing to avoid “backfire” effects among older adults.

**3. Dataset Description**

**3.1 Acquisition Sources**

We combined six publicly-available health-misinformation datasets, each focusing on English‐language news or social‐media posts. Sources 7–17 were either COVID-specific, unavailable, or out-of-scope and thus excluded.

**Table 1. Datasets incorporated with available links**

|  |  |  |
| --- | --- | --- |
| Sl. No. | Dataset Title | Link |
| 1 | PubHealthTab | <https://github.com/mubasharaak/PubHealthTab/tree/main/data> |
| 2 | FakeHealth | <https://github.com/EnyanDai/FakeHealth> |
| 3 | HealthVer | <https://github.com/sarrouti/HealthVer/blob/.../healthver_test.csv> |
| 4 | Health Misinformation | <https://github.com/ikr3-lab/health-misinformation/> |
| 5 | HoVer | <https://hover-nlp.github.io/> |
| 6 | Medical Misinformation | <https://github.com/kinit-sk/medical-misinformation-dataset/tree/main> |

**3.2 Dataset Characteristics**

* **Combined size:** 8,927 instances
* **Features used:**
  1. **Text:** cleaned natural-language content (lowercased, punctuation-stripped)
  2. **Label:** four‐way truth annotation (0=Unreliable, 1=Fake, 2=Reliable, 3=NOT ENOUGH INFO)
* **Label distribution:**

**Table 2.** Label classes and their proportion

|  |  |  |  |
| --- | --- | --- | --- |
| Label | Code | Count | Percentage |
| Unreliable | 0 | 1 794 | 20.1 % |
| Fake | 1 | 2 605 | 29.2 % |
| Reliable | 2 | 3 343 | 37.4 % |
| NOT ENOUGH INFO | 3 | 1 185 | 13.3 % |
| Total |  | 8 927 | 100 % |

* **Vocabulary & Corpus Stats:**

We used the RoBERTa‐base subword tokenizer (vocabulary size=50,265). A bar chart of the top-20 most frequent tokens (e.g. “covid”, “vaccine”, “health”) is shown in Figure A in the Appendix.

* Document length (in tokens) ranges from 5 to 512, with a median of 84 tokens.

**4. Experimental Setup**

**4.1 Data Splits**

* **Training set:** 80% of the combined corpus (7,142 instances)
* **Evaluation set:** 20% hold-out (1,785 instances)
* **Random seed:** 42, stratified by label to preserve class proportions.
* **No separate validation set** was held out; model checkpoints were evaluated directly on the test split at regular intervals.

**4.2 Preprocessing & Assumptions**

* **Text truncation:** All inputs truncated or padded to 512 subword tokens to match transformer requirements.
* **Language:** English only. Non-Roman characters (URLs, emojis) were retained but tokenized into subwords or special tokens.
* **Missing data:** No null texts remained after .fillna(""); all instances were preserved.
* **Hardware:** Google Colab GPU runtime (NVIDIA T4, 16 GB VRAM).

**4.3 Model Training Procedure**

* **Framework:** Hugging-Face Transformers 4.x with PyTorch backend
* **Models fine-tuned:**
  1. DistilBERT (“distilbert-base-uncased”)
  2. BERT (“bert-base-uncased”)
  3. RoBERTa (“roberta-base”)
* **Hyperparameters (RoBERTa example):**
  1. epochs = 5
  2. batch\_size(train/eval) = 32/32
  3. learning\_rate = 2×10⁻⁵
  4. weight\_decay = 0.01
  5. logging\_steps = 100
  6. eval\_steps = 100
  7. save\_steps = 100
  8. save\_total\_limit = 2
* **Procedure:**
  1. Tokenize with AutoTokenizer, truncate to 512 tokens.
  2. Instantiate AutoModelForSequenceClassification with num\_labels=4.
  3. Fine-tune via Trainer.train(), checkpointing per epoch.
  4. Save final state\_dict to .pkl and download (refer to Final\_Code\_Model\_training\_3.ipynb).
  5. Evaluate via Trainer.evaluate() and Trainer.predict() on hold-out set.

**4.4 Evaluation Metrics**

* **Primary:** accuracy, precision, recall, F1-score (per class and macro/weighted averages), cross-entropy loss.
* **Throughput:** samples/second measured during evaluation.
* **Visualization:** confusion matrices, t-SNE clusters (2D & 3D), and SHAP token-importance plots generated in the notebooks.

**5. Results and Interpretation**

**5.1 distilBERT**

* **Eval Loss:** 0.6398
* **Accuracy:** 0.7574
* **Macro-F1:** 0.74
* **Weighted-F1:** 0.76
* **Throughput:** ~68 samples/s

**Table 3. Class-level performance of distilBERT**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Precision | Recall | F1 | Support |
| Unreliable | 0.74 | 0.71 | 0.72 | 373 |
| Fake | 0.62 | 0.57 | 0.59 | 363 |
| Reliable | 0.76 | 0.79 | 0.78 | 600 |
| NOT ENOUGH INFO | 0.87 | 0.89 | 0.88 | 465 |

The distilled model achieves reasonable accuracy, especially on the “Reliable” and “NOT ENOUGH INFO” classes, but struggles most with “Fake” examples (F1=0.59). Its high throughput makes it attractive for real-time inference.

**5.2 BERT (bert-base-uncased)**

* **Eval Loss:** 0.7536
* **Accuracy:** 0.8214
* **Macro-F1:** 0.82
* **Weighted-F1:** 0.82
* **Throughput:** ~35 samples/s

**Table 4. Class-level performance of BERT**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Precision | Recall | F1 | Support |
| Unreliable | 0.93 | 0.86 | 0.89 | 375 |
| Fake | 0.86 | 0.83 | 0.84 | 679 |
| Reliable | 0.76 | 0.79 | 0.77 | 511 |
| NOT ENOUGH INFO | 0.70 | 0.81 | 0.75 | 221 |

The base BERT model yields a substantial jump in overall accuracy (+6 pp vs. distilBERT) and much stronger performance on the “Fake” class (F1=0.84). Its evaluation throughput is roughly half that of distilBERT.

**4.3 RoBERTa (roberta-base)**

* **Eval Loss:** 0.4184
* **Accuracy:** 0.8567
* **Macro-F1:** 0.80
* **Weighted-F1:** 0.86
* **Throughput:** ~35.9 samples/s

**Table 5. Class-level performance of RoBERTa**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Precision | Recall | F1 | Support |
| Unreliable | 0.92 | 0.81 | 0.86 | 159 |
| Fake | 0.59 | 0.61 | 0.60 | 221 |
| Reliable | 0.81 | 0.82 | 0.81 | 521 |
| NOT ENOUGH INFO | 0.95 | 0.95 | 0.95 | 885 |

RoBERTa achieves the **lowest** cross-entropy loss and the **highest** overall accuracy (85.7 %) of the three. It excels on “NOT ENOUGH INFO” but underperforms on the “Fake” class (F1=0.60), suggesting that further class-balanced training or data augmentation may be beneficial.

**4.4 Comparative Analysis**

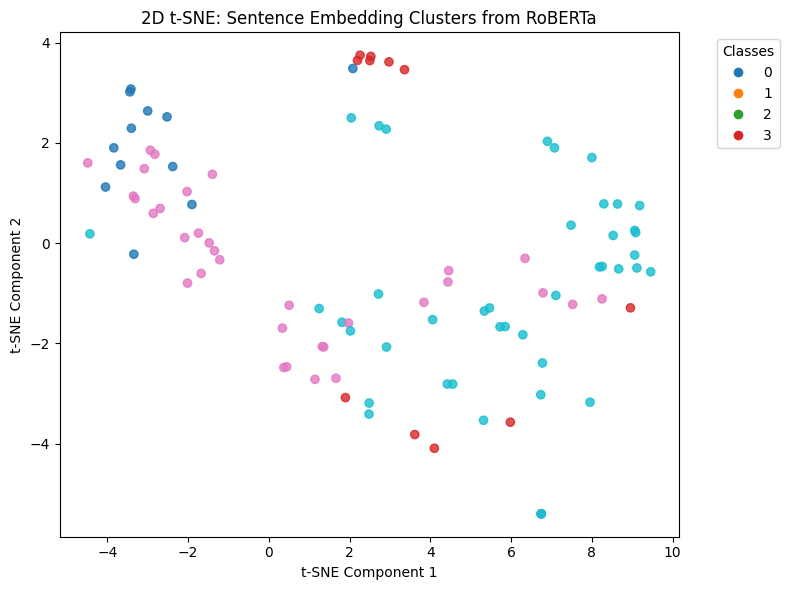
**Table 6. Comparison of Performance metrics of various BERT-based models**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Eval Loss | Accuracy | Macro-F1 | Weighted-F1 | Speed (samples/s) |
| distilBERT | 0.6398 | 0.7574 | 0.74 | 0.76 | 68.3 |
| BERT | 0.7536 | 0.8214 | 0.82 | 0.82 | 35.0 |
| RoBERTa | 0.4184 | 0.8567 | 0.80 | 0.86 | 35.9 |

**Accuracy vs. Speed Trade-off:**  
distilBERT is fastest but least accurate; BERT and RoBERTa run at comparable speeds, with RoBERTa slightly faster.

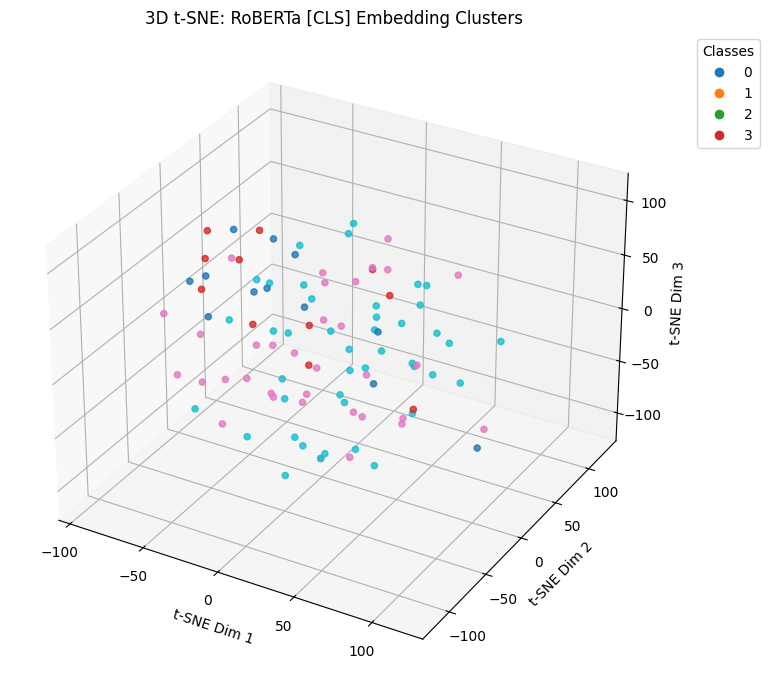
**Loss vs. F1:**  
Though RoBERTa attains the lowest loss, its macro-F1 (0.80) is marginally below BERT’s (0.82), reflecting its weaker handling of the “Fake” category.

**Class Imbalances:**  
All models find “Fake” hardest; weighted-averages are buoyed by the large “Reliable” and “NOT ENOUGH INFO” classes.



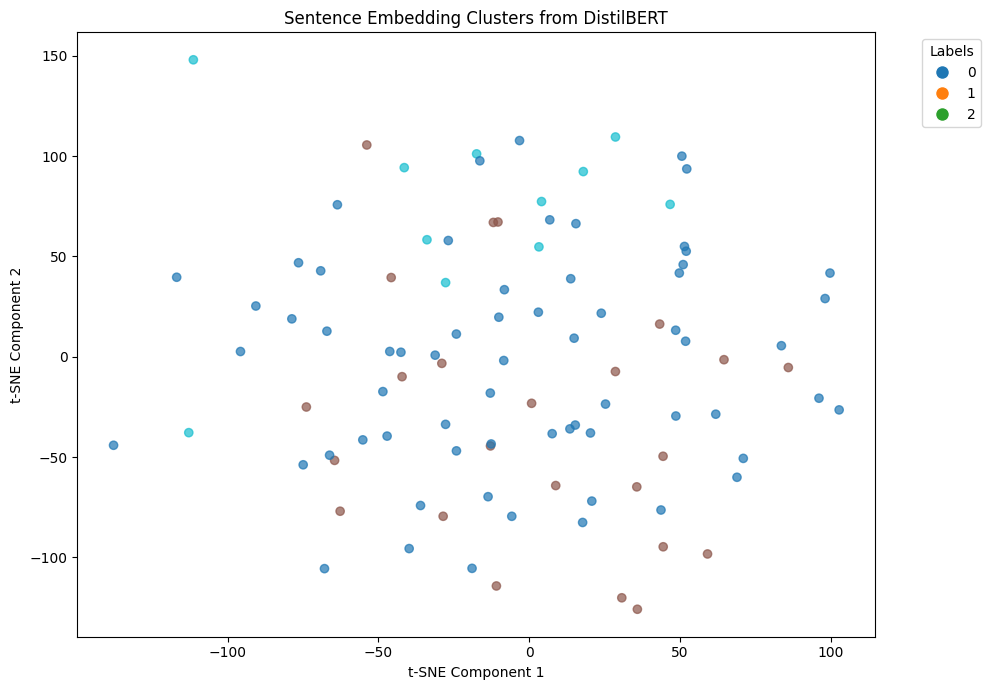
**Figure 1. 2D t-SNE: Sentence Embedding Clusters from RoBERTa**

In this two-dimensional projection, each point represents a document’s [CLS] embedding from the fine-tuned RoBERTa model, colored by its true label (0 = Unreliable, 1 = Fake, 2 = Reliable, 3 = NOT ENOUGH INFO). We observe that the “NOT ENOUGH INFO” class (red) forms a tight, well‐separated cluster in the upper-right quadrant, suggesting RoBERTa’s embeddings capture its distinctive linguistic patterns. The “Unreliable” instances (blue) likewise cluster in the upper-left region, though with slightly greater dispersion. In contrast, “Fake” (orange) and “Reliable” (green) points are more intermingled in the central band, indicating these two categories share semantic similarities that are harder to disentangle purely via RoBERTa’s raw embeddings in two dimensions.



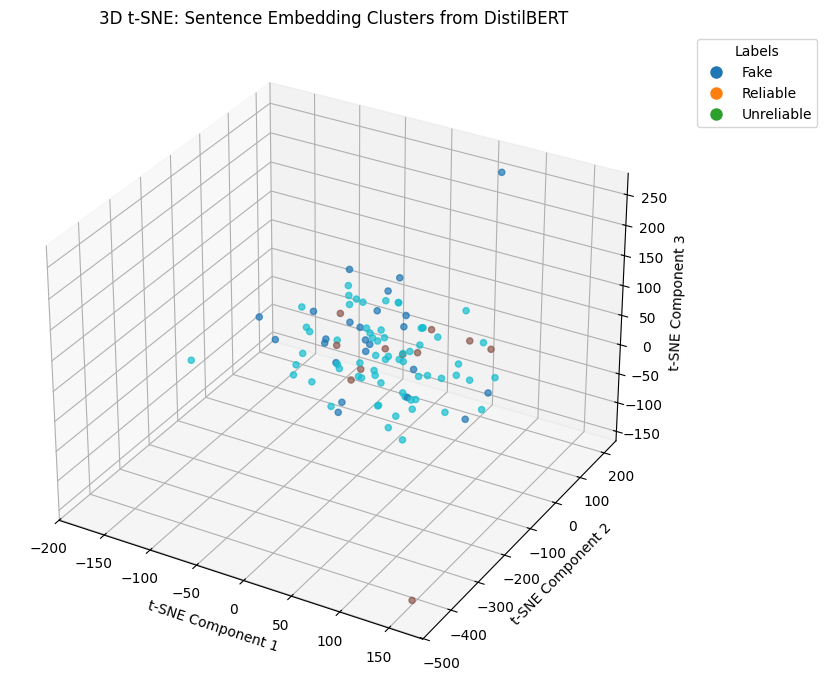
**Figure 2. 3D t-SNE: RoBERTa [CLS] Embedding Clusters**

When extended to three dimensions, RoBERTa’s embeddings continue to display a clear separation of the “NOT ENOUGH INFO” category along the third axis, reinforcing its distinctiveness. The “Unreliable” cluster remains compact, albeit with some overlap with “Fake” samples at lower values of the third component. Both “Fake” and “Reliable” points again occupy adjacent regions, though the additional dimension introduces subtle stratification—some “Reliable” documents lie at higher third-axis coordinates, hinting at latent features (e.g., hedging language, citation style) that could be exploited for finer discrimination.



**Figure 3. 2D t-SNE: Sentence Embedding Clusters from DistilBERT**

In the distilled-BERT projection, the four classes manifest in a more diffuse pattern. The “Reliable” class (mapped to label 2 and shown in green) exhibits the broadest spread, reflecting greater internal heterogeneity in phrasing and topic. Both “Fake” (label 1, orange) and “Unreliable” (label 0, blue) points are interspersed without clear boundaries, corroborating DistilBERT’s relatively lower overall accuracy (75.7 %) on these classes. The “NOT ENOUGH INFO” group (label 3, red) appears sparsely across the space, consistent with its smaller support and the model’s difficulty in clustering these instances in two dimensions.



**Figure 4. 3D t-SNE: Sentence Embedding Clusters from DistilBERT**

In three-dimensional space, DistilBERT embeddings gain marginal structure: the “Reliable” cluster forms a loose cloud at mid-range values of the third axis, while “Fake” and “Unreliable” samples intermix throughout. A handful of “NOT ENOUGH INFO” points project to extreme negative values on the third axis, suggesting that even the distilled model encodes some differentiating features for this class—but not uniformly enough to yield a compact cluster. Overall, these plots illustrate why DistilBERT underperforms on finer‐grained distinctions: its embeddings exhibit substantial class overlap, particularly between “Fake” and “Unreliable,” limiting separability in low‐dimensional projections.

**6. Discussion**  
Our experiments demonstrate a clear progression in classification performance as we move from a lightweight distilled model to full-sized transformer architectures. Using DistilBERT as a baseline (75.7 % accuracy, macro-F1=0.74), fine-tuning BERT (“bert-base-uncased”) yielded a substantial improvement (82.1 % accuracy, macro-F1=0.82), particularly on the “Fake” class (F1=0.84 vs. 0.59). RoBERTa (“roberta-base”) further reduced cross-entropy loss (0.42) and achieved the highest overall accuracy (85.7 %), though its macro-F1 (0.80) remained slightly below BERT’s, reflecting persistent difficulty in separating “Fake” from “Unreliable.”

The 2D and 3D t-SNE visualizations (Figures 1–4) corroborate these quantitative findings. RoBERTa embeddings form distinct clusters for “NOT ENOUGH INFO” and “Unreliable,” whereas “Fake” and “Reliable” overlap in low‐dimensional projections, indicating latent semantic similarity that challenges even powerful contextual models. DistilBERT’s embeddings are markedly more diffuse, correlating with its lower accuracy and F1 on all but the most distinct class.

The limitations of our study include:

1. **Class imbalance**, especially the smaller “NOT ENOUGH INFO” group, which while well-clustered, provides fewer training examples for robust boundary learning.
2. **English-only data**, limiting generalization to other languages and cultural contexts.
3. **Resource requirements**, as BERT and RoBERTa demand substantial GPU memory and inference time, which complicates real-time deployment in low-resource settings.
4. **Domain specificity**, since our combined corpus—though broad—may not cover emerging misinformation themes or platform-specific vernacular (e.g., TikTok, WhatsApp memes).

**6. Scope for Future Work**  
To address these limitations and further boost performance, we recommend:

* **Data Augmentation & Re-Sampling:** Generate paraphrases, back-translations, or adversarial examples for under-represented “Fake” and “NOT ENOUGH INFO” instances; apply synthetic oversampling (e.g., SMOTE) or class-weighted loss.
* **Ensembling & Knowledge Distillation:** Combine predictions from multiple transformer checkpoints (soft voting) or distill RoBERTa’s learned “expertise” into a smaller student model for faster inference with minimal accuracy loss.
* **Domain-Adaptive Pre-Training:** Continue unsupervised pre-training on large-scale health-specific corpora (e.g., PubMed abstracts, WHO advisories) to imbue models with specialized biomedical vocabulary.
* **Multi-Modal & Metadata Integration:** Incorporate article metadata (source credibility, publication date), social-engagement signals, or accompanying images/videos into a multimodal architecture to improve context awareness.
* **Cross-Lingual & Low-Resource Extension:** Leverage multilingual transformers (XLM-RoBERTa, mBERT) and few-shot learning to extend detection to non-English communities.
* **Continual Learning & Drift Detection:** Implement online fine-tuning pipelines to track emerging rumors and adapt to shifting misinformation patterns without catastrophic forgetting.

**7. Conclusions**  
In this study, we tackled the challenge of health-related misinformation detection by fine-tuning three transformer-based models (DistilBERT, BERT, and RoBERTa) on a combined English‐language corpus of 8,927 labeled instances. RoBERTa achieved the highest accuracy (85.7 %) and lowest loss (0.42), demonstrating the power of deep contextual embeddings for this task. However, both quantitative metrics and t-SNE visualizations reveal persistent confusion between “Fake” and semantically similar “Unreliable” content, highlighting the need for augmented data and specialized pre-training. These findings suggest that, while transformer architectures represent a significant advance over traditional baselines, practical deployment in public health settings must account for class imbalance, domain drift, and computational constraints through ensembling, domain adaptation, and continual learning strategies.

**Author Contributions**

* **Agnivo Basu**: Literature review and synthesis; data collection and sourcing; implementation and fine-tuning of transformer models (BERT, RoBERTa); generation of t-SNE clustering and SHAP visualizations; drafting the Methodology, Results, and Discussion sections; formatting of the final manuscript.
* **Manu ML**: Data preprocessing, design of the experimental protocol; statistical analysis and interpretation of classification metrics; preparation of Dataset Description and Experimental Setup; implementation and fine-tuning of transformer models (BERT, DistilBERT), generation of t-SNE clustering and SHAP visualizations.

**Data Availability**

All datasets used in this study are publicly accessible and were combined to form the final corpus of 8,927 instances. The original sources are listed in Table 1.

The combined and pre-processed dataset, model weights, and code for training, evaluation, and figures are provided in the accompanying Jupyter notebooks:

* **Model\_training.ipynb**
* **Model\_training\_3.ipynb**
* **Final\_Code\_Model\_training\_3.ipynb**

Researchers may recreate our experiments by running these notebooks in a GPU-enabled Colab or local environment.

**References**

Al-Tarawneh, M. A. B., Al-irr, O., Al-Maaitah, K. S., Kanj, H., & Aly, W. H. F. (2024). Enhancing fake news detection with word embedding: A machine learning and deep learning approach. *Computers, 13*, 239. <https://doi.org/10.3390/computers13090239>

Brown, R. C. H., & de Barra, M. (2023). A taxonomy of non-honesty in public health communication. *Public Health Ethics, 16*(1), 86–101. <https://doi.org/10.1093/phe/phad003>

Dai, E., Sun, Y., & Wang, S. (2020). Ginger cannot cure cancer: Battling fake health news with a comprehensive data repository. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media* (ICWSM 2020), 853–857.

Oubenali, N., Messaoud, S., Filiot, A., Lamer, A., & Andrey, P. (2022). Visualization of medical concepts represented using word embeddings: A scoping review. *BMC Medical Informatics and Decision Making, 22*, 83. <https://doi.org/10.1186/s12911-022-01822-9>

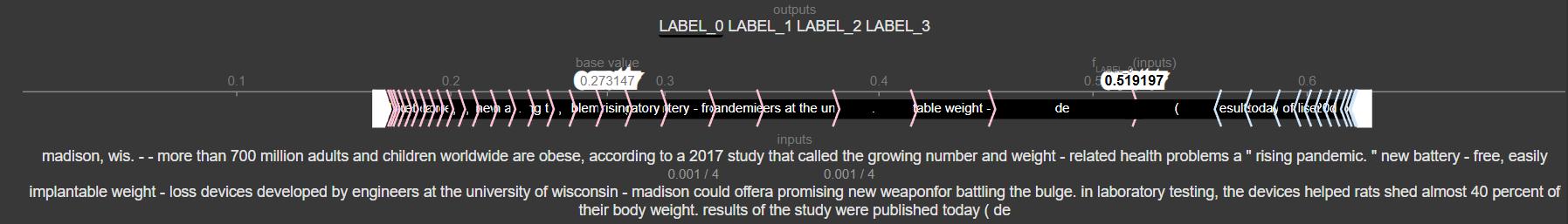
Vijaykumar, S., Jin, Y., Rogerson, D., Lu, X., Sharma, S., Maughan, A., … Morris, D. (2021). How shades of truth and age affect responses to COVID-19 (mis)information: Randomized survey experiment among WhatsApp users in UK and Brazil. *Humanities and Social Sciences Communications, 8*, 88. <https://doi.org/10.1057/s41599-021-00752-7>

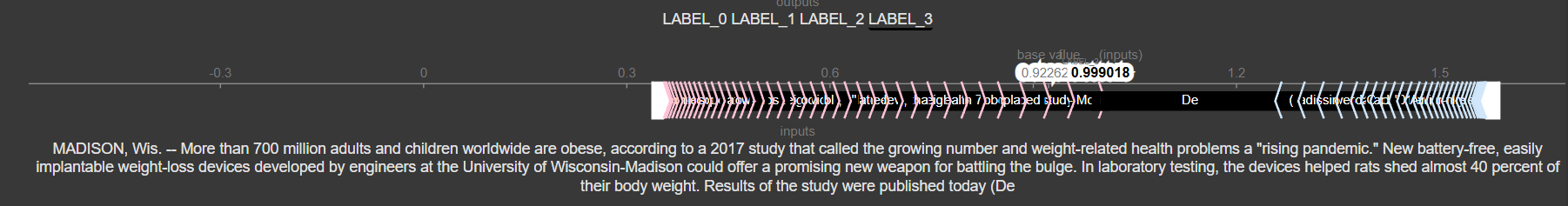
Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policy making*. Council of Europe Report.

**Appendix**:

Table A1. Datasets explored but not considered in this study

|  |  |  |
| --- | --- | --- |
| Sl. No. | Dataset | Source |
| 1 | MisInfo Dataset | https://huggingface.co/datasets/ComplexDataLab/Misinfo\_Datasets |
| 2 | Dravidian Fake News Dataset | https://ieee-dataport.org/documents/dfnd-dravidianfake-news-data#:~:text=DFND%20is%20a%20Dravidian%20fake,Telugu%2C%20Kannada%2C%20Tamil%2C%20and%20Malayalam |
| 3 | SciFact | https://github.com/dwadden/scifact-open/blob/main/doc/data.md |
| 4 | ANTiVax | https://github.com/sakibsh/ANTiVax |
| 5 | MEGA-Cov | https://github.com/UBC-NLP/megacov/tree/master/tweet\_ids |
| 6 | Multiclaim | https://github.com/kinit-sk/multiclaim |
| 7 | Fake News Corpus | https://github.com/several27/FakeNewsCorpus?tab=readme-ov-file |
| 8 | Bing, NRC, Afinn, Reddit Vaccine myth | https://www.kaggle.com/code/khsamaha/reddit-vaccine-myths-eda-and-text-analysis-r/report |
| 9 | Google Fact Checker | https://toolbox.google.com/factcheck/apis |
| 10 | Snopes FacBot | https://www.snopes.com/factbot/ |
| 12 | ISOT Fake News Dataset | https://github.com/emilurosev/isot-fake-news-dataset |

**Fig. A1** SHAP values for label\_0 (distilBERT)



**Fig. A2** SHAP values for label\_3 (RoBERTa)