# DiaBERT: Combating Diabetes Misinformation Using Transformer-Based Models

## Project Overview

DiaBERT is an end-to-end misinformation detection system tailored to diabetes-related content. Built on the BioBERT transformer model and enhanced through domain adaptation (DANN), DiaBERT classifies online health claims into **True**, **False**, or **Partially True**. It is deployed as a **Chrome extension** that provides real-time credibility classification and explanation for users encountering health-related content online.

## Key Features

* **Transformer Backbone**: Built on BioBERT (a BERT model pre-trained on biomedical corpora)
* **Domain Adaptation**: Implemented DANN (Domain-Adversarial Neural Network) to adapt from formal (medical) to informal (social media) domains
* **Three-Class Classification**: True, False, Partially True
* **Content Filtering**: SBERT + Cosine Similarity to filter only diabetes-related input
* **Explainability**: Worked with LIME, SHAP, and Transformers Interpret with final choice being Transformers Interpret due to alignment with transformer architecture
* **Deployment**: Real-time Chrome Extension using ONNX-optimized BioBERT model via Flask API hosted on Fly.io

## Dataset

1. **Formal Dataset**:
   * Derived from the DETERRENT dataset
   * 2269 diabetes-related claims (True: 1661, False: 608)
2. **Informal Dataset**:
   * Curated from Facebook, Twitter (X), and Reddit
   * Manually annotated into 3 classes (True, False, Partially True)
   * 902 diabetes-related claims (True:575, False:167, Partially True:160)
   * Texts underwent preprocessing: normalization, unicode correction, contraction expansion, emoji/URL filtering

## Model Pipeline

1. **Stage 1**: Supervised fine-tuning of BioBERT on formal two-class data
2. **Stage 2**: Domain Adaptation using DANN — encoder learns invariant features between formal and informal domains
3. **Stage 3**: Final supervised fine-tuning on informal three-class data

## Content Filtering (SBERT + Cosine Similarity)

* SBERT model: all-MiniLM-L6-v2
* Averaged embedding vector created from diabetes domain corpus
* Queries must pass a cosine similarity threshold (> 0.7) to be considered “in-domain”

## Explainability

* **Tried**: SHAP, LIME, Transformers Interpret
* **Chosen**: **Transformers Interpret** (attention-based saliency, integrated gradients)
  + Highlights tokens contributing to prediction
  + Easily integrates into the transformer pipeline
  + Works well with subword tokenization

## Deployment

* **Backend**: Flask API
  + Predict endpoint: /predict
  + Health check: /ping
  + Includes SBERT filtering and explanation generation (GPT or template)
  + Deployed via **Fly.io**, optimized with **ONNX** for faster inference
* **Frontend**: Chrome Extension
  + User selects or enters text
  + Calls /predict endpoint
  + Receives classification + explanation
  + Feedback rendered within browser popup

## Sample Use Case

User visits a blog post that claims: *“bitter leaf cures diabetes completely.”*

* DiaBERT classifies it as **False**
* Highlights tokens: *“cures”, “completely”*
* Explanation: “The claim suggests a definitive cure without scientific support. Bitter leaf may help regulate blood sugar but is not a standalone treatment.”

## Resources

* Extension: https://chromewebstore.google.com/detail/diabert-classifier/pkccflhgplpbmoglflfjhhlnpdjbblpk?authuser=0&hl=en-GB
* Code, datasets, and training scripts

## Future Work

* Expand dataset to include **non-English** and **multilingual** claims
* Extend DiaBERT to cover other chronic illnesses (e.g., asthma, hypertension)
* Evaluate model bias and build user-centric explanation toggles

## License

This project is released under the MIT License.