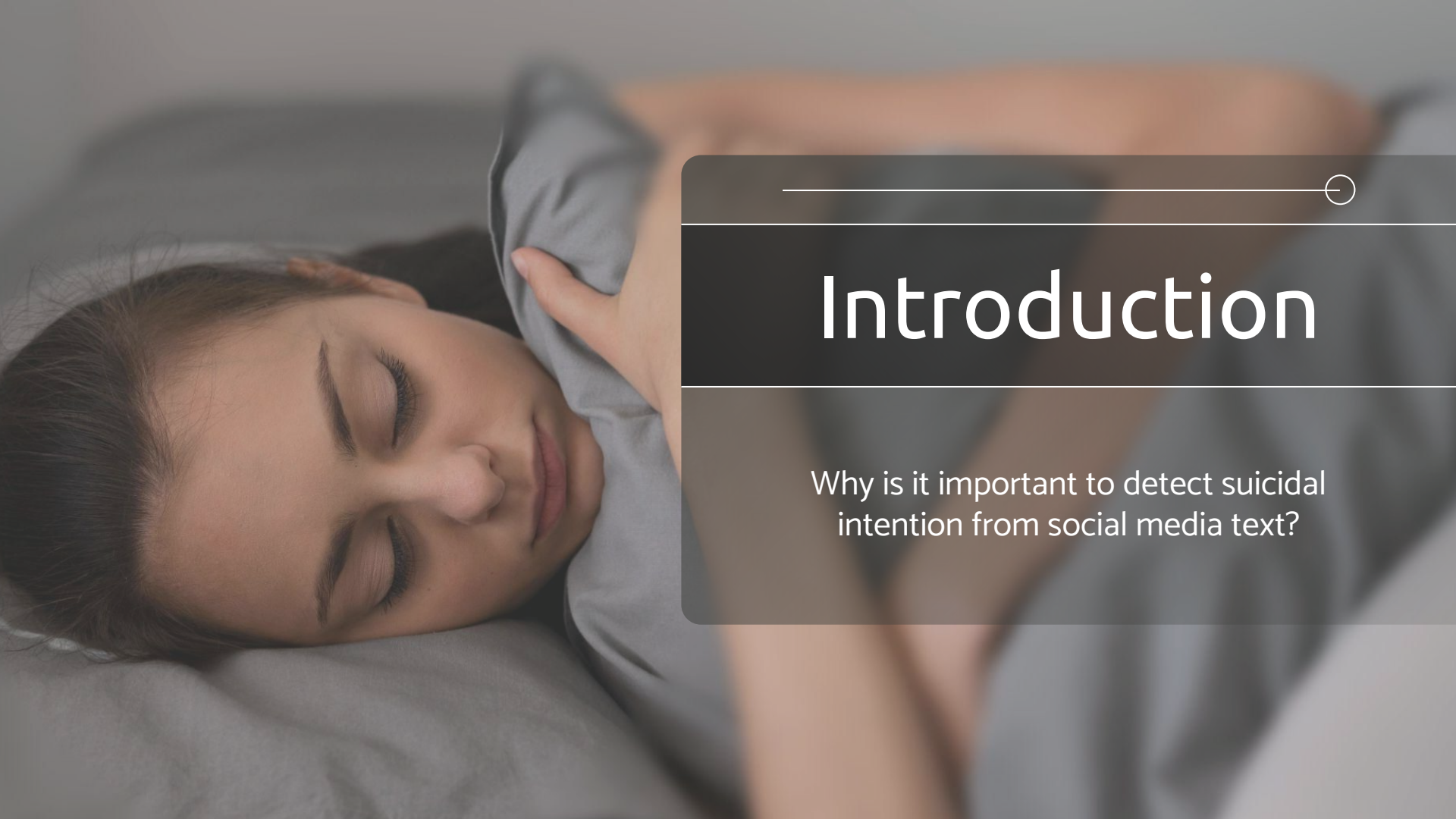
Suicidal Text Detection

Apply BERT Language Model to Identify Suicidal Texts

Chih-Han (Maggie) Chuang May, 2023

Code: https://github.com/ChuangChihHan/Suicidal-Text-Detection/blob/master/Suicidal_Detection_Bert_Maggie_Chuang.ipynb



Introduction

Why is it important to detect suicidal intention from social media text?

Suicide represents a significant social issue

- More than 700 000 people die due to suicide every year.
- Suicide is the second leading cause of death among people aged between 10 and 34 years
- Due to the stigma surrounding medical treatments, individuals with suicidal thoughts often avoid seeking help. Instead, they tend to express their intentions to commit suicide through social media platforms.
- Because mental illness may be diagnosed and treated, the early identification of warning signs may be the most effective way of preventing suicide.

Table of contents

01. Opportunity Statement

The benefits of detecting suicidal intention from social media text

Goal & Opportunity Statement

Developing BERT language models capable of detecting early signs of suicidal ideation in social media posts can benefit mental health industry tremendously.

Early Intervention

Enable timely intervention and prevent self-harm or loss of life

Improved Mental Health Support

Offer more personalized support to individuals in need, enhancing the effectiveness of the mental health services.

Resource Optimization

Prioritizing individuals at a higher risk based on the analysis of social media data

Scalability

Efficiently process a large volume of social media posts, making it scalable for businesses to analyze a wide range of user-generated content across various platforms.

02. Data Overview

Data introduction, assumption, NLP data preprocessing techniques

Data Introduction

	Text	Class
0	Am I weird I don't get affected by compliments if it's coming from someone I know irl but I feel really good when internet strangers do it	non-suicide
1	Honetly idkI dont know what im even doing here. I just feel like there is nothing and nowhere for me. All i can feel is either nothing or unbearably sad. Im ignoring friends every opitunity i can. I feel like im loosing my girlfriend. I only hurt everyone i talk too and i dont cause anything good. Im behind on my education, i feel alone but for the first time its not a feeling ive enjoyed. I have no hopes or dreams. I care about nothing, not family, not friends, not even my girlfriend (i still love her, its complicated and i dont have the words to describe it). I would do something to end myself but i know im not strong and brave enough to do it, and knowing im that weak makes me sadder. The only thing i can do is push away all emotion and be empty, because as bad as it is im used to it, its my way of being normal.	suicide

Data Source	a collection of posts from ”SuicideWatch” and ”depression” subreddits of the Reddit platform.
Data Volume	232,074 rows, 2 columns
Variables	Text: post content Class: suicide / non-suicide
Temporal Coverage	SuicideWatch: Dec 16, 2008 - Jan 2, 2021 Depression: Jan 1, 2009 - Jan 2, 2021

Data Assumption

1	2	3	4
Data Authenticity	User Identification	Data Labeling Accuracy	Language Consistency
Posts are genuine and accurately represent the thoughts of the users posted them	Posts in the dataset are from identified users of the "SuicideWatch" and "depression" subreddits	The posts in the dataset have been accurately labeled to indicate whether they contain suicide thoughts	The posts in the dataset are all in accurate modern English

Data Cleaning & Preprocessing

Lower Casing	Special characters Removal	Stopword removal	Tokenization	Data Shuffling
Convert all the text into lower case to reduce the dimensionality of the input space and aid in generalization	Remove punctuations, numerical data, multiple whitespaces, duplicate characters	Remove low information words from the text so can focus on important ones	Use BERT tokenizer to incorporate special tokens and attention masks	Randomizing the order of the training sets to prevent any inherent order or patterns in data

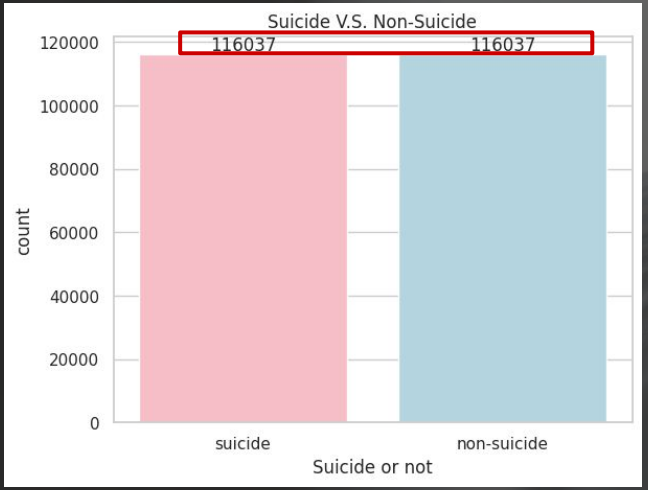
03.

EDA & Feature Engineering

Create new variables, plot message length distribution, token visualization

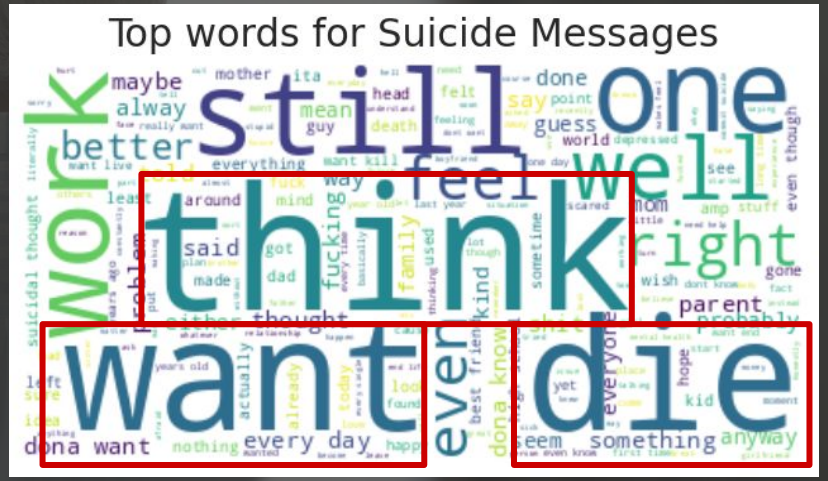
Exploratory Data Analysis

Class Distribution



Suicide and non-suicide messages have the same amount (116,037) → Balanced dataset

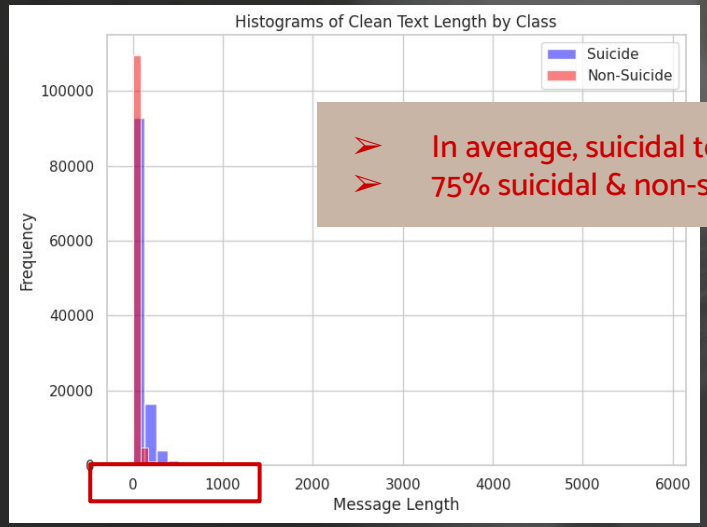
Top words in suicidal messages



The most frequent words in suicidal messages: think, still, want, die, work

Feature Engineering

Message Length histogram by class



- In average, suicidal text length > non-suicidal
- 75% suicidal & non-suicidal text lengh < 500 words

	Suicide		Non-suicide	
	Message length	Clean text length	Message length	Clean text length
mean	202	90	60	29
			154	73
			2	1
25%	60	27	18	9
50%	127	57	31	16
75%	250	111	60	29
max	9684	3874	14632	5850

- Create message length variable to count the number of words in each sentence
- Create label variable to map suicide → 1 ; non-suicide → 0 for text classification usage

04. Model Analysis

Fine Tune pre-trained BERT model
Overfitting & Regularization

Reasons of using BERT on classifying suicidal text

1	2	3	4
Transfer Learning	Contextual Understanding	Pre-trained knowledge	Capturing Relationships
Learns from a large dataset and can be fine-tuned on smaller task-specific datasets, saving computational resources and time.	Captures the contextual meaning of words by considering the surrounding words. Deeper understanding of the text → improved classification accuracy.	Is pre-trained on unlabeled data from BookCorpus and Wikipedia. Able to leverage this knowledge on detecting suicidal text.	The attention mechanism allows BERT to assign different weights or attention scores to each word based on its relevance and importance to other words in the sequence.

How BERT works on classifying suicidal text

Input: Features

Text

Am I weird I don't get affected by compliments

→

BERT

Train the entire architecture

→

Classifier

Sigmoid function

→

Output

1: Suicidal

0: Non-Suicidal

Prepare the input data for BERT model

- Tokenize sentences using BERTtokenizer
  - padding → all the messages have the same length
  - special tokens → [CLS] (classification token) & [SEP] (separator token)
  - Get input_ids & attention masks needed for the model
- Save the processed data into pickle files so don't need to do encodings repeatedly & load the pickle files

Set up pre-trained BERT Model: TFBertForSequenceClassification

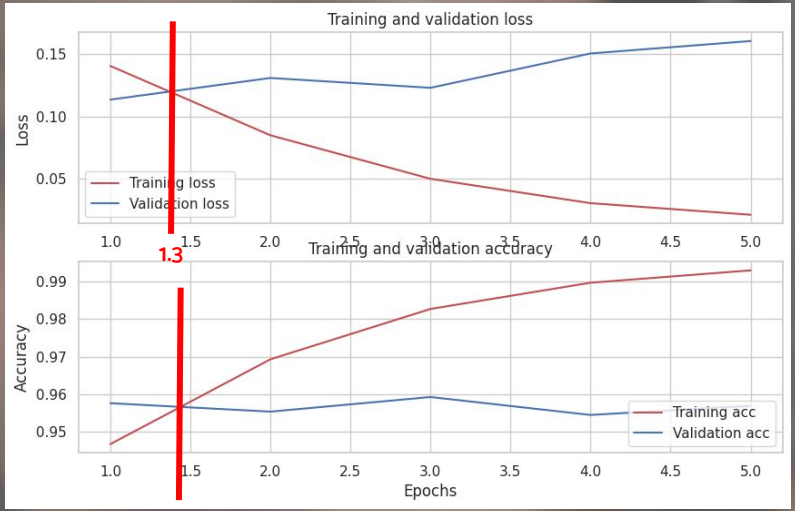
- Split data into training data(80%), testing data(20%)
- Set up the loss, metric, and optimizer and compile the model
- Split data into training (75%) & validation data (25%) and fit the compiled model on training data
- Monitor training loss, acc & validation loss, acc to check if overfitting/ underfitting
- Fine tune the model

Fine tuning BERT - Model 1

Model 1

- Batch size = 32
- Epochs = 5
- Loss: Sparse Categorical Cross entropy
- Optimizer: Adam
  - Learning rate: 2e - 5
  - epsilon=1e-08
- Metrics: Sparse Categorical Accuracy

Overfitting



- **Adam:**
  optimizes the update step for each parameter by incorporating momentum and adaptive learning rates

- Validation loss > training loss at epoch = 1.3
- Next step: <u>smaller epoch + regularization to address overfitting</u>

Fine tuning BERT - Model 2 (regularization)

Model 2

- Batch size = 32
- Epochs = 1
- Loss: Sparse Categorical Cross entropy
- Optimizer: **AdamW**
  - Learning rate: 2e - 5
  - **Weight decay = 0.01**
- Metrics: Sparse Categorical Accuracy

Regularization

AdamW

- an extension of the Adam optimizer that includes **L2 weight decay** during parameter updates.
- L2 weight decay (weight regularization) **adds a penalty term** to the loss function.
- Including weight decay → **prevent overfitting** by encouraging smaller weights & **improve generalization**.

- Training Loss: 0.1413 | Training_acc: 0.9470
- Val_loss: 0.1141 | val_accuracy: 0.9577

→ Train_acc ≈ Val_acc → **Resolve overfitting**

05. Results & Findings

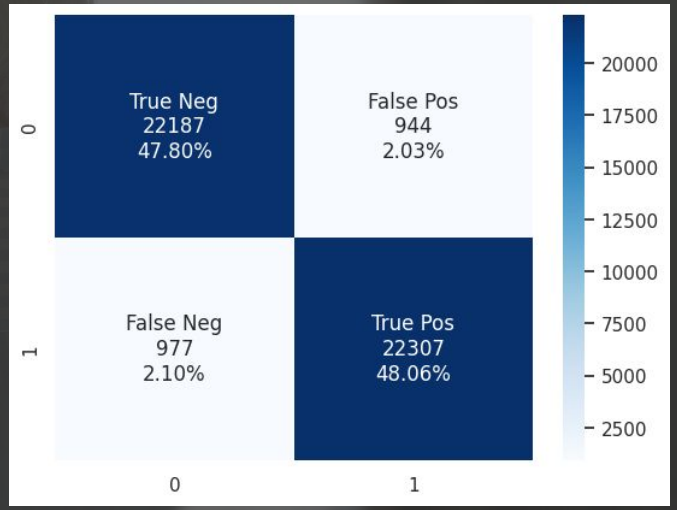
Model Evaluations & Predictions
Learnings

Model Evaluation - Model 1 Overfit

Classification Report

	precision	recall	f1-score	support
0	0.50	0.49	0.49	23179
1	0.50	0.51	0.50	23236
accuracy			0.50	46415
macro avg	0.50	0.50	0.50	46415
weighted avg	0.50	0.50	0.50	46415

Confusion Matrix



➢ **Precision, recall, accuracy are inconsistent with confusion matrix** → might be due to **overfitting**

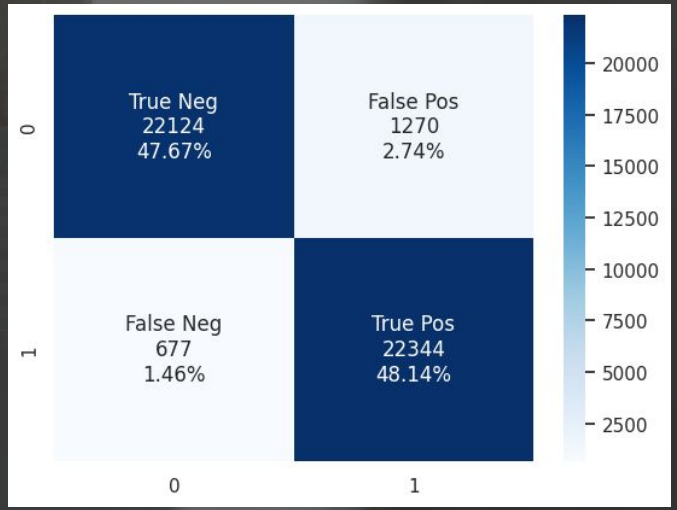
True Negative + True Positive ≈ 96%

Model Evaluation - Model 2 (Regularization)

Classification Report

	precision	recall	f1-score	support
0	0.97	0.95	0.96	23394
1	0.95	0.97	0.96	23021
accuracy			0.96	46415
macro avg	0.96	0.96	0.96	46415
weighted avg	0.96	0.96	0.96	46415

Confusion Matrix



➢ **Model 2's performance is better** than Model 1 → use model 2 to do prediction

True Negative + True Positive ≈ 96%

Model 2 Prediction

Input Text (Unseen data)	Output
Today I felt good in the morning, everything was good, but in the evening, it rained, and as a result, I got stuck in traffic. My life sucks; I should end it; I should kill myself.	1: Suicide
Today, I felt good in the morning; everything was good, but in the evening, it rained, and as a result, I got stuck in traffic.	0: Non-Suicide

Fit model on test dataset → Test Accuracy: 95.79% || Try out new messages, the results are all correct

Learnings

- Data Preprocessing:
  - For classification tasks, remember to check if the classes are balanced.
  - For NLP tasks, data cleaning & tokenization & embedding are crucial steps before feeding into BERT.
  - Store the processed data into pickle files so don't need to do encodings repeatedly
- Modeling:
  - BERT model learns from a large dataset and can be fine-tuned on smaller task-specific datasets. This saves computational resources and time.
- Evaluation
  - Monitor training loss, acc & validation loss, acc to check if overfitting/ underfitting occurs for fine-tuning models
- Prediction
  - Evaluate the model on unseen data(split the data into train, validation & test data or create new input text as test data) to check model performance

06. Future Work

Model Improvement

Transfer Learning & Domain Adaptation

Leverage pre-trained models or embeddings specifically trained on mental health or suicide-related data to enhance the model's understanding of the domain-specific language and context.

Ensemble Methods

Combine multiple BERT models with different configurations or ensemble them with other models (e.g., traditional machine learning classifiers or other deep learning models) to leverage their complementary strengths.

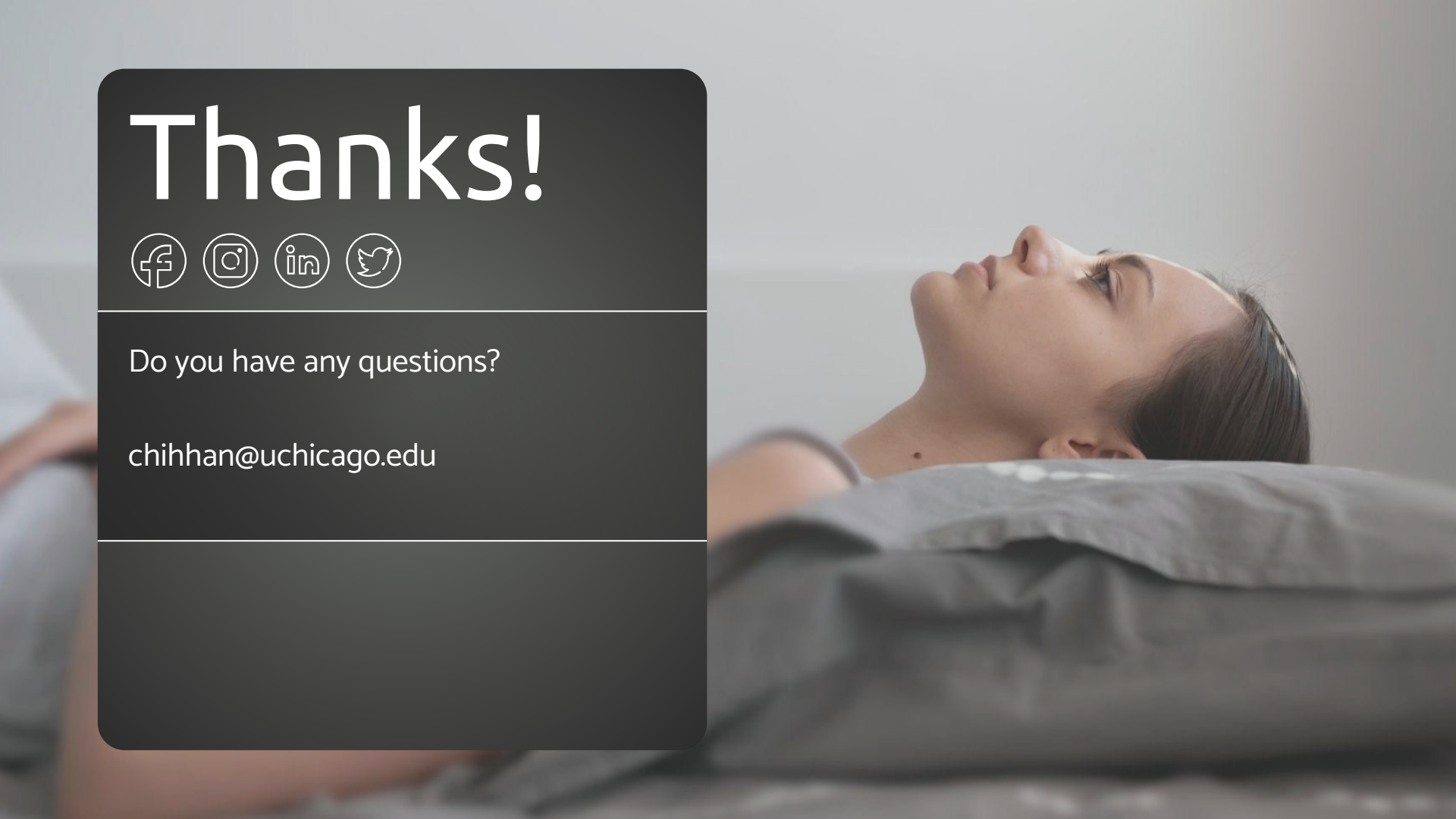
Error Analysis & Feedback Loop

Continuously analyze the model's predictions and collect feedback from domain experts to identify patterns of misclassification. This feedback can help refine the model, update training data, or adjust classification thresholds.

Thanks!

Do you have any questions?

chihhan@uchicago.edu



References

- WHO-*Suicide:* https://www.who.int/news-room/fact-sheets/detail/suicide
- Detecting and Analyzing Suicidal Ideation on Social Media Using Deep Learning and Machine Learning Models: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9565132/
- Classify text with BERT: https://www.tensorflow.org/text/tutorials/classify_text_with_bert
- NLP for Suicide and Depression Identification with Noisy Labels: https://towardsdatascience.com/nlp-for-suicide-and-depression-identification-with-noisy-labels-98d7bb98f3e8
- Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing: https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html
- Spam Email Classification using BERT: https://www.kaggle.com/code/kshitij192/spam-email-classification-using-bert
- BERT Text Classification using Keras: https://swatimeena989.medium.com/bert-text-classification-using-keras-903671e0207d
- BERT: https://huggingface.co/docs/transformers/model_doc/bert