Fighting Health-Related Misinformation in Social Media With Large Language Models

By

Moisés Robles Pagán

A thesis submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE
in
COMPUTER ENGINEERING

UNIVERSITY OF PUERTO RICO
MAYAGÜEZ CAMPUS
2024

Approved by:

Manuel Rodríguez Martínez, Ph.D. President, Graduate Committee	Date
Emmanuel Arzuaga, Ph.D. Member, Graduate Committee	Date
Domingo Rodríguez Rodríguez, Ph.D. Member, Graduate Committee	Date
FirstName I. LastName, Ph.D. Representative of Graduate Studies	Date
José Cedeño Maldonado, Ph.D. Department Chairperson	Date

ABSTRACT

Hi! We encourage you to visit https://libguides.uprm.edu/writingclinics and check out the **Abstracts Clinic.** Keep in mind that depending on your discipline, abstracts should be a **single paragraph**, containing no more than **150 words** for theses or **350 words** for dissertations. It should concisely but clearly summarize your thesis document. The **IMRaD format** is recommended for writing abstracts: Introduction (1-3 sentences long, present tense), Methodology (1-3 sentences long, past tense), Results (1-3 sentences long, past tense), and Discussion (1-2 sentences long, present tense). Remember that the number of sentences and verb tense are only guidelines!

RESUMEN

El Resumen debe ser una traduccion del Abstract. No deben diferir en contenido.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

ACKNOWLEDGMENTS

I want to thank the GRIC personnel! :D

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Contents

List of Figures

List of Tables

List of Abbreviation

AI	Artificial Intelligence
API	Application Program Interface
CLM	Causal Language Modeling
CPU	Central Processing Unit
CUDA	Compute Unified Device Architecture
DBMS	Database Management System
GPU	Graphic Processing Unit
HF	Hugging Face
JSON	JavaScript Object Notation
LLM	Large Language Model
LoRA	Low-Rank Adaptation
LSTM	Long Short-Term Memory
MLM	Mask Language Modeling
NIH	National Institutes of Health
NLP	Natural Language Processing
PEFT	Parameter Efficient Fine-Tuning
PMC	PubMed Central
RNN	Recurrent Neural Network
SQL	Structured Query Language
THS	Twitter Health Surveillance

UPRM University of Puerto Rico Mayagüez

List of Appendices

Chapter 1

Introduction

1.1 Motivation

This project was created because

Example of numbered items. The work is divided in three phases:

1. Collect data
2. Build model
3. Validate results

1.2 Objectives

The objectives of this project are as follows:

- Identify and extract information from official health sources: This data will be stored in a vector database that the model will use as context to rebut the misinformation. The model will cite official health sources related to the tweets to sustain their classification.

- Identify and finetune a Large Language Model: Select an appropriate base Large Language Model architecture that will:

  1. Detect if a text is health related.
  2. Determine if a text is misinformation.
  3. Use official health sources texts to combat the texts classified as misinformation and cite from the gather data.

- Compare with the previous version of THS: To measure the effectiveness of the classification with the LLM, we are going to compare it with the previous THS results and validate the advantages of a Large Language Model on solving Natural Language Processing problems.

1.3 Contributions

- **Finetune Large Language Models for health classification on social media:** Large Language Models are being used for different fields nowadays. However, these do not focus on health misinformation on social medias. We present Large Language Models as a solution to classify and rebut health misinformation texts on social medias, and use research papers extracted from PubMed as context for the LLM.

-

- **Pending Contribution Title:** We used 12,441 texts for the health-related classification labeled as related, unrelated, or ambiguous. For the misinformation-classification we had 8,772 texts labeled as misinformation or not misinformation. For the model rebuttal, we extracted 56,365 papers from PubMed.

1.4 Outline

This paper has the following organization. Chapter 2 contains the literature review on Transformers, Large Language Models, the different use cases of these models for classification, and misinformation on social media. Additionally, we describe the importance of disinformation on social medias. For Chapter 3, we can observe the problem description and methodology. Later, on Chapter 4, we have the experiment, and the projects pipeline for the training and classification. Chapter 5 presents our results based on accuracy and performance. In Chapter 6, related works are presented, with our conclusion and suggestions or future work.

Chapter 2

Literature Review

2.1 Introduction

There has been many advancement....

2.2 Large Language Models (LLM)

Natural language processing (NLP) has always been an intricate field because of the complexity of how humans communicate. The meaning of a message can vary because of homonyms, tone, context, and other factors that affect the message delivered; these are some challenges computers face when trying to replicate or learn text communication and expressions. However, this changed with the introduction of Large Language Models (LLM) [1]. These models are trained with large amounts of data to replicate human-like patterns or generate text based on statistical relationships between words, and all was made possible because of transformers [2]. Previous NLP techniques such as Recurrent Neural Network (RNN) and Long-Short Term Memory (LSTM) could understand a sentence context in the short term. However, these struggle when trying to understand longer texts. In contrast, transformers architecture, seen in Figure 2.1, differs from others because it uses self-attention.

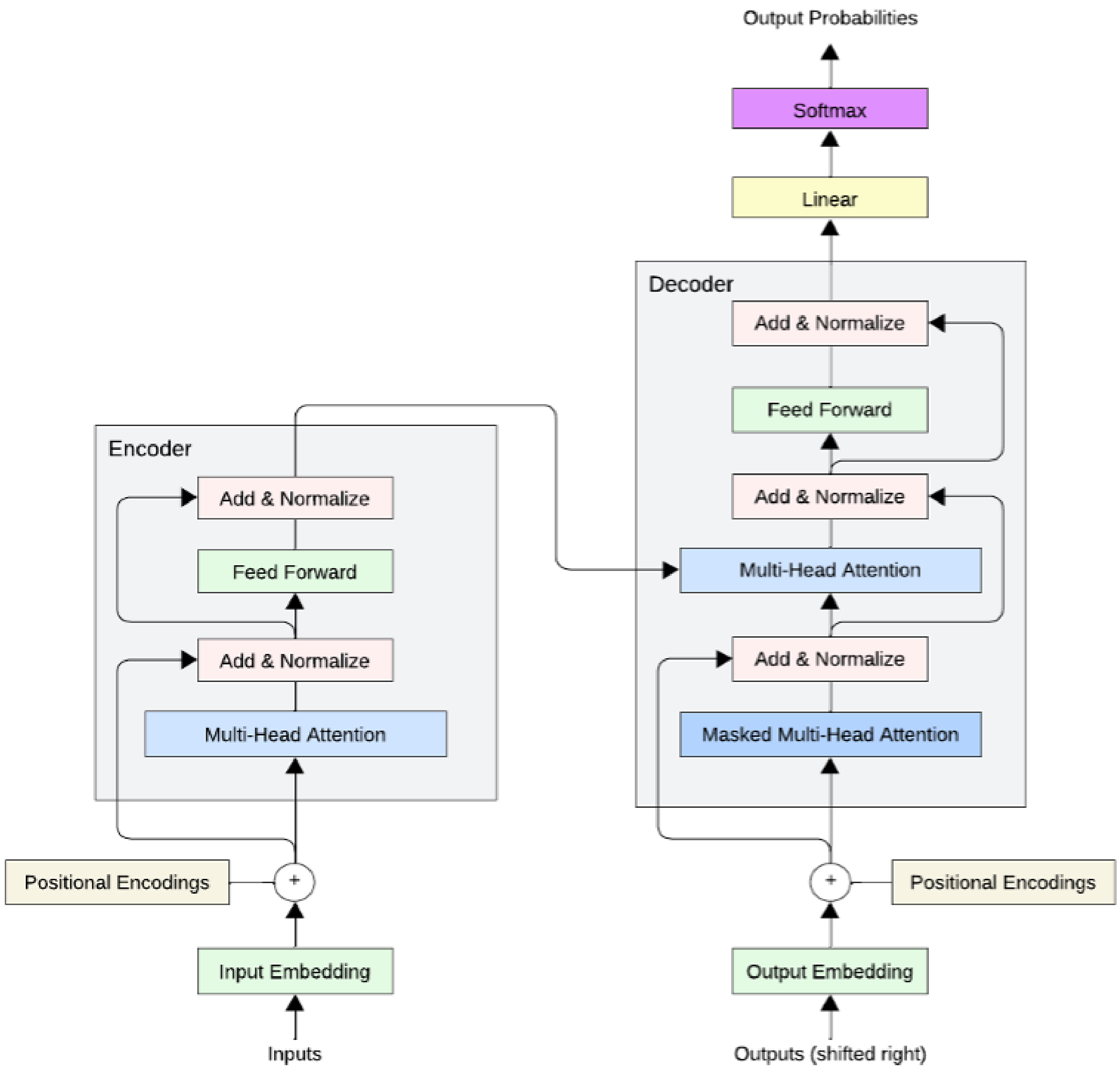


Figure 2.1: The Transformer Architecture

This self-attention finds dependencies between all words in a text, short and long-term. The process of this starts by turning words into tokens. A token can be a word, subword, individual letter, or a sequence of words mapped to an embedding. The embedding is the vector representation of a token in high-dimension space; its size depends on how much information it stores about the token. Now, self-attention finds relationships between all tokens and gives them an attention score, measuring the relevance of a token to others. Transformers uses the scores to generate a final representation of each token. This process depends on how the models make a token. The tokenization strategy is determined by the preprocessing stage, and influenced by the embedding and model architecture. The embedding impacts the strategy because of its dimensionality, the amount of information encoded, and the model's sequence length limitations. In Figure 2.1, we can see an encoder and a decoder in the transformers architecture. Based on that, there are different LLM architectures: decoder-only, encoder-only, and encoder-decoder. Each one has advantages for specific tasks and limitations for others.

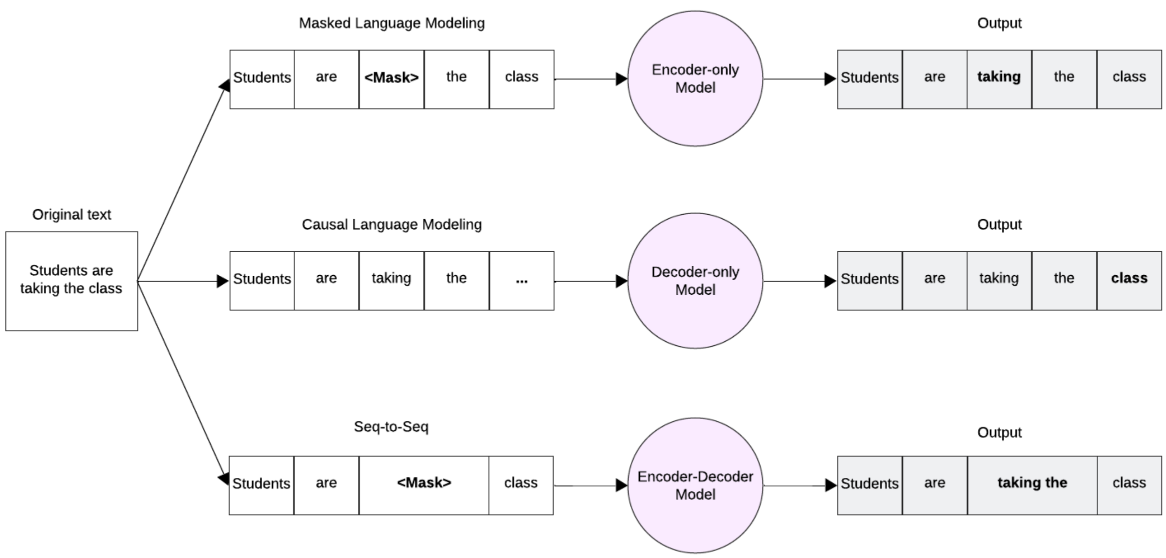


Figure 2.2: LLM Architecture and Comparison

2.2.1 Decoder-only models

Decoder-only models predict the next word based on the previous context, thus being unidirectional models. The model achieves this by taking a text or prompt as input and returning a first word. For each subsequent word, the model uses the previously generated text as input to predict the next word, continuing until it produces a coherent output. In Figure 2.2, the model predicts the last word based the previous context.

Because of their ability to predict sequences of texts, they are frequently employed in tasks like summarization and text generation. Models that perform those task are also called causal language modeling (CLM), because they predict new tokens not found in the input. Compared to the other architectures, these models are massive in size. Because of their sizes, these are not very practical or cost-effective for daily usage. These are the most commonly known models, such as GPT-3 [3], Mistral [4], and LLaMa [5].

2.2.2 Encoder-only models

This type of model predicts by masking specific words in a sentence. Said masking helps them understand the meaning or relation of the masked word based on context. They are bi-directional, which means they take the context before and after the masking to evaluate the word. The example in Figure 2.2 shows a sentence with a word mask and being inputted to the encoder LLM; the model predicts the missing word by using the surrounding text. That is why they tend to perform well at classification and sentiment analysis but are not optimal for text generation. Said model are called masked language models (MLM). In contrast to other architectures, these models are relatively small. Some examples of encoder-only LLM are BERT [6] and RoBERTa [7].

2.2.3 Encoder-Decoder models

Encoder-decoder models combines masking and text generation. The way the work is by masking sequence of texts and using the context around it to make a prediction. As seen in Figure 2.2, they can mask more than one word from the original inputted text. Because the model generates sequences of texts, it is commonly used for translation and question and answer. It is know that translating from one language to another is not possible using a word by word translation; one must understand the entire sequence to not loss context. To have an optimal model, both encoder and decoder must be trained for the task one wishes to achieve. Depending on the task, it can be harder to train compared to the other types of architectures. Bart [8] and T5 [9] are example of encoder-decoder LLMs.

2.2.4 Classification tasks

Large Language Models use different learning methods to train. Such methods include zero-shot, one-shot, and few-shot learning. As the name implies, zero-shot learning is training a model without previous knowledge of the data; it is learning from scratch. One-shot learning is a method that receives one example as input and tries to generalize from that example. The final method uses multiple examples to find a pattern between them.

In Figure [**IMAGE**], we can see an example of zero-shot learning. This LLM has no previous knowledge of the task that it must perform. Nonetheless, the model used, GPT-3, has the advantage that it can follow instructions when redacted clearly. Here, we tell the model that it must act as a medical expert and identify if a message is health-related and why it is classified that way. Also, the result must follow a specific format. This process of instructions is called prompt engineering. Prompting does not retrain or adjust the model parameters. Thus, it does not always have optimal results.

[**FIGURE**]

Moreover, LLMs that completed training with no additional modifications are known as based models. The response from this model will most likely make no sense with the premises it receives. This happens because the model is trained on excessive texts to find patterns between them, but this pattern might not be on par with the input. For the model to return a coherent response, it must go through a finetuning process. Finetuning consists of making a model perform specific tasks, such as chatting, summarization, chatbots, and others. To finetune a model, it must undergo another training process, but now the training data relates to the tasks it will perform.

The authors in [**REFERENCE**] created a model to understand images and give a text description or a combination of text and image. The model was trained on images and captions of those images using zero-shot. Their resulting model was used to create a chatbot that identifies images, answers questions, or gives visual examples about them. Another experiment was [**REFERENCE**], where the authors trained a model to identify sentiments on financial market decisions. The advantage of in-context learning of the model resulted in a 70% accuracy on sentiment prediction, failing mostly on neutral posts. That experiment showed the problems that users face when they use social media to make decisions. They clarified the importance of not taking the model for granted and how social media can cause a user to make a poor decision.

2.3 Misinformation in Social Media

There are many sources in the world to find information about any topic. Nonetheless, many people use social media as their primary source and mostly take this information as truth without validation. On occasion, these can be fake, misleading, or wrong. When this happens unintentionally or by lack of understanding of the topic, it is called misinformation. On the other hand, when it is intentional to provide wrong information, this is known as disinformation. For simplification, both terms will be used interchangeably.

Misinformation has been dangerous during critical events like natural disasters or health crises. For instance, during the COVID-19 pandemic, falsehood arose saying that the vaccine had microchips, which led many people to die [1] because they refused to get vaccinated out of fear. A problem with disinformation is that the audience does not always detect it. When misinformation spreads and is not clarified early on, it can be confused as fact. Older audiences and people with less education are more likely to share and believe fake or misleading news [**REFERENCE**].

There has been various research on reducing the propagation of misinformation. Misinformation was seen as a game-theoretic in [**REFERENCE**], where some players spread fake news, and others tried to stop it. They created an agent at the network level to combat misinformation in a simulation. However, they could not conclude the efficiency of their model because of the lack of a pattern. On the other hand, in [**REFERENCE**] the authors used LSTM and BERT to classify misinformation from the different news sources. They proved that BERT outperformed LSTM, achieving an accuracy of 64.88% against 60.59%. Another approach to detecting fake information was on [**REFERENCE**], where they detected fake LinkedIn profiles. On this occasion, the dataset used for training included real and AI-generated profiles. They tested multiple LLMs, but BERT resulted in the highest accuracy of 95.67%. These investigations prove the efficiency of Large Language Models for Natural Language Processing in the misinformation field. Regardless, none of these studies addresses health-related misinformation on social media.

Chapter 3

Citations, Images, and Equations

You can visit the Overleaf Documentation library at: https://www.overleaf.com/learn or the official LaTeX wiki at: **https://en.wikibooks.org/wiki/LaTeX** for in-depth guides to the LaTeX typesetting system.

3.1 How to incorporate citations in your document

Make sure you have your references file in **.bib format**. You can export it from Mendeley and copy/paste them into the **referencias bib** file. Using the **citecommand**, start writing your source identifier and Overleaf will automatically show you the available references. A sample citation [10]. If you want to cite two references [11, 12] or [11],[12]. Citing a range of references [11–13].

3.1.1 Citation Styles

There are several bibliography styles to choose from: **abbrv, acm, alpha, apalike, ieeetr, plain, siam, unsrt.** The default is ieeetr. Remember to change the option in the document preamble in **tesis.tex**

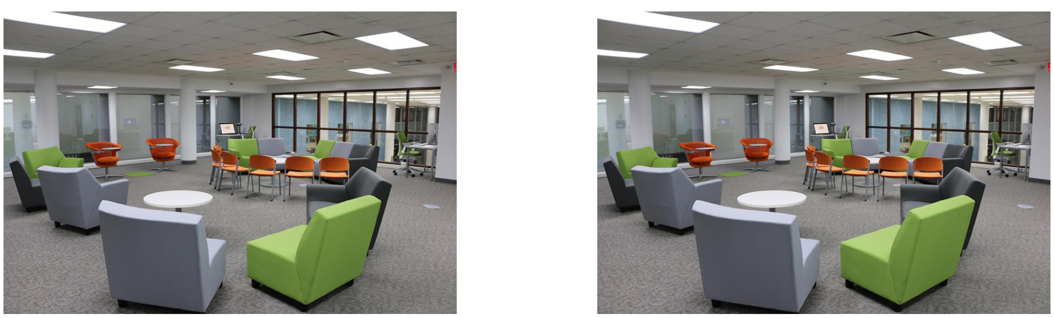
3.2 Using Images

This is an example of a figure, as shown in Figure 3.1. Your images must be uploaded to the **images** folder. Accepted file formats are **pdf, png, jpg, and eps**. Use the following options for the location of the image on the page: [**htbp**] that refer to here, top, bottom, or special page. Pay attention to the image width. If the image is wider than the margins, LaTeX will produce an warning or error.

Figure 3.1: GRIC logo!!

3.2.1 Adding Subfigures

If you need to place two subfigures in your figure, follow the example below:



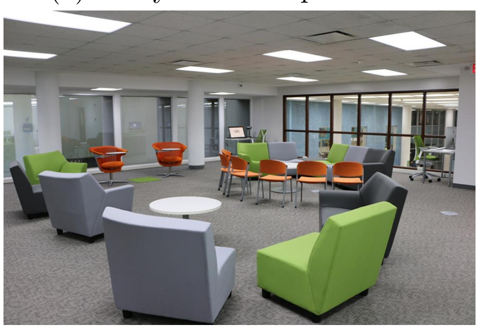
(a) Put your sub-caption here d (b) Put your sub-caption here

Figure 3.2: Put your caption here

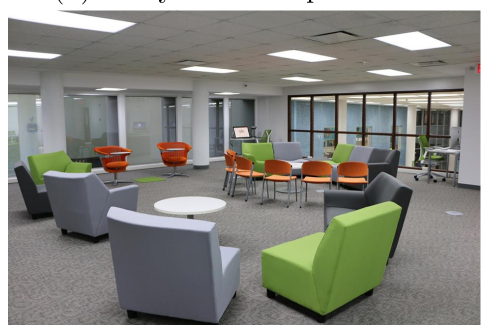
If you need to place four subfigures in your figure, follow the example below, but LaTeX is very particular with widths, so you have to play around with the numbers.It might give you overflow warnings. These won't stop your document from compiling. It is easier to build the four images as a single one before uploading it to Overleaf.

(a) Put your sub-caption here

(b) Put your sub-caption here



(c) Put your sub-caption here



(d) Put your sub-caption here

Figure 3.3: Put your caption here

Chapter 4

System Architecture

4.1 System Overview

4.2 Fine-tuning

The Large Language Models (LLM) used for this paper are pre-trained, meaning that the model learned how words relate to each other, the model is trained to understand a language. This process is computationally expensive and requires a large amount of data. Instead of creating a model from scratch, we can teach an existing model to learn a new task such as text classification, translation, generation, or other. This is called fine-tuning, and in this case, it was used to teach the model to classify two types of texts: health-related and misinformation-related.

The health-related dataset comprises over 12,441 tweets extracted from the previous THS project. Said dataset was classified into three categories: related, unrelated, and ambiguous tweets. A second dataset, with 8,762 texts, was created with data from different sources such as news, social media, and blogs classified as misinformation and non-misinformation ***[Insert References]***.

The models used for the classification process were Bert, T5, and LLaMa-2.

4.2.1 Health-Related Classification

4.2.2 Misinformation Classification

2. Misinformation classification

LR, Batch size, seed, and epochs are static.

- Each model was fine-tune twice.

1. Sequence classification: 1, 2, or 3; 1 or 2.

2. Classification with text generation: Related, Unrelated, or Ambiguous; Misinformation or Not Misinformation.

- Weighted average added for the sequence classification.

4.3 Paper ETL Pipeline

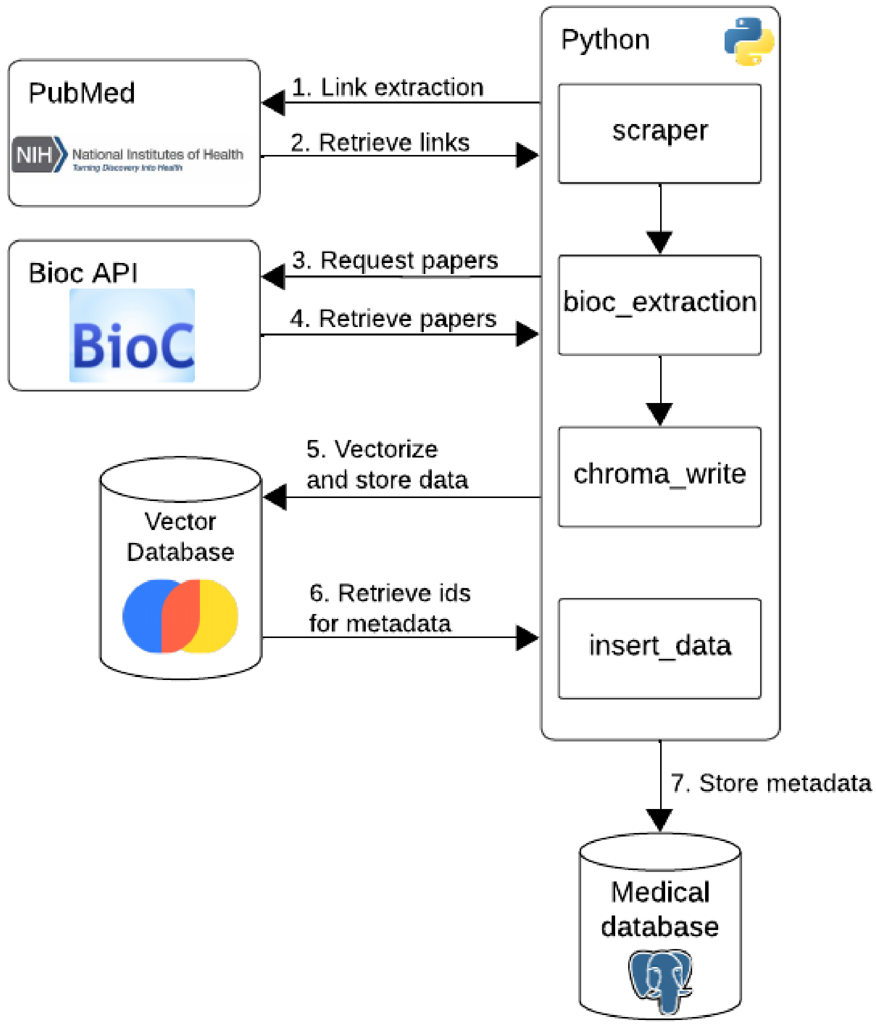


Figure 4.1: Medical Data Extraction Pipeline

For our models to be able to rebut the misinformation, it needs to use credible sources. We identified PubMed to extract the papers and stored them in a vector database. To extract these papers, we used the BioC API [14] that contains the PubMed papers. However, the API needs the research papers' identifiers. To get these identifiers a scraper was designed to extract the PubMed Central (PMC) identifier. The pipeline in The pipeline in Figure 4.1 goes as follows:

1. Extract papers identifier: We selected 14 different keywords for the papers that would be extracted. For each keyword, we retrieved 5,000 PMC identifiers and stored them in CSV files.

2. Paper requests: After retrieving the identifier, the system made requests to the API and the results were saved locally in JSON format. Each JSON was preprocessed to only contain text.

3. Vectorizing data: The research paper's contexts were broken into chunks, vectorized by an LLM **REFERENCES**, and stored in a chroma database. Each chunk was given a unique identifier to be paired with the original text.

4. Storing metadata: With all papers vectorized, the paper's metadata and the chunk's unique identifiers were stored in a Postgres database. Duplicate records and researchs that had their reference missing were removed, to prevent inconsistency and ensure that our classifier cites the correct sources. Our relational database Table diagram can be found on **ADD TABLE DIAGRAM REFERENCE**

4.4 Misinformation Rebuttal Pipeline

The pipeline shown in Figure 4.2 goes as follows:

1. Health-related classification: Verifies if the inputted text is health-related. The possible options are related, unrelated, or ambiguous.

2. Misinformation classification: Checks if the text contains misinformation. If any misinformation is detected, we need to find official health data to rebut said misinformation.

3. Context finder: A query is created for a vector database based on the original text. This query is sent to a vectorized database, chroma [reference].

4. Medical information database: This is a database that contains medical data from official sources. Returns the IDs and strings of texts that are related to the original text.

5. Organize and rebut: The result from the medical database is now processed and used to make a rebuttal for the misinformation in the text. Then, we query the relational database to extract the references of the papers used for the previous part. The output includes the original text, the health and misinformation classifications, the correction of the misinformation, and the citation of the sources used.

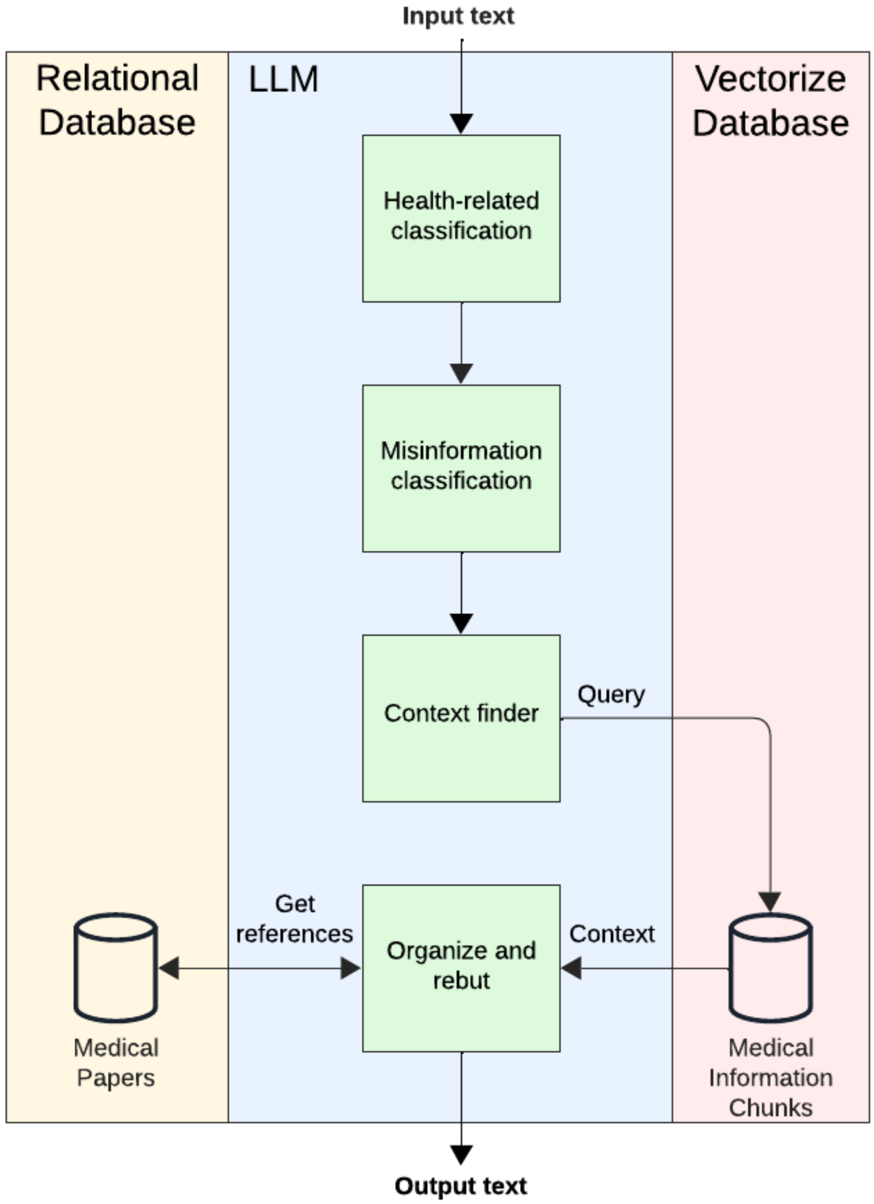


Figure 4.2: Misinformation Rebuttal LLM System Architecture

4.5 Hardware and Software

- V100 machines: 32Gb VRAM, and 80-ish RAM

  Cuda 11.7

- Python 3.9.19
- Pytorch 2.0.1
- Transformers 4.34.0

Chapter 5

Literature Review

For tips and guidelines on how to write your Literature Review, visit the GWF clinic section at: https://libguides.uprm.edu/writingclinics, Clinics 2020.

5.1 Section

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris.

5.1.1 Subsection

Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque.

5.1.1.1 Subsubsection

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

5.1.2 Subsection

Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque.

5.1.2.1 Subsubsection

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque.

5.2 Section

Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem.

5.2.1 Subsection

Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla.

Chapter 6

Methodology

6.1 Section

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris.

6.1.1 Subsection

Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque.

6.1.1.1 Subsubsection

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

6.1.2 Subsection

Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque.

Chapter 7

Results

7.1 Section

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris.

7.1.1 Subsection

Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque.

7.1.1.1 Subsubsection

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

7.1.2 Subsection

Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque.

7.1.2.1 Subsubsection

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque.

7.2 Section

Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem.

7.2.1 Subsection

Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla.

Chapter 8

Conclusions

8.1 Section

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque.

8.1.1 Subsection

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque.

8.1.2 Subsection

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque.

Bibliography

[1] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, and A. Mian, “A comprehensive overview of large language models,” 2024.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2023.

[3] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, “Language models are few-shot learners,” *CoRR*, vol. abs/2005.14165, 2020.

[4] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, “Mistral 7b,” 2023.

[5] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin,

E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," 2023.

[6] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018.

[7] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019.

[8] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension," 2019.

[9] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei, "Scaling instruction-finetuned language models," 2022.

[10] P. A. M. Dirac, *The Principles of Quantum Mechanics*. International series of monographs on physics, Clarendon Press, 1981.

[11] A. Einstein, "Zur Elektrodynamik bewegter Körper. (German) [On the electrodynamics of moving bodies]," *Annalen der Physik*, vol. 322, no. 10, pp. 891–921, 1905.

[12] D. E. Knuth, *Fundamental Algorithms*, ch. 1.2. Addison-Wesley, 1973.

[13] A. López and K. Soderstrom, "Insolation in Puerto Rico," *Journal of Solar Energy Engineering*, 1983.

[14] D. C. Comeau, C.-H. Wei, R. Islamaj Doğan, and Z. Lu, "PMC text mining subset in BioC: about three million full-text articles and growing," *Bioinformatics*, vol. 35, pp. 3533–3535, 01 2019.

Appendix A: MATLAB Code

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Appendix B: Data

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris.

Appendix C: More Data

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris.

Appendix D: More Data

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris.