AI-Generated Content Detection

Candidate Number: 277191

Supervisor: Julie Weeds

MSc Data Science, University of Sussex

May 2024

Contents

Abstract 3

1. Introduction 3

2. Background Research 4

    2.1 Motivation 4

    2.2 Literature Review 6

        2.2.1 Team1 - LLMLab 6

        2.2.2 Team2 - Nicholas Broad 7

3. Dataset and Methodology 7

    3.1 Dataset 7

    3.2 Methodology 8

4. Project Milestones 13

Bibliography 15

Abstract

Many large language models (LLMs) like OpenAI's ChatGPT ("ChatGPT," 2022) or Claude3 ("Claude," 2023) have evolved rapidly. It has been a challenging task to distinguish between AI-generated content from human-written text. Countless potential problems are emerging due to the rapid development of these LLMs, especially in the education field. Students are being accused of AI plagiarism (Nguyen, 2023) . Teachers can't tell if the content is genuine or AI-generated. This phenomenon has impacted both learning and teaching greatly. This research aims to first analyze award-winning models from the Kaggle competition "LLM - Detect AI-Generated Text." ("LLM - Detect AI Generated Text," 2024). We then used different LLMs to generate AI essays to balance the size of the given dataset. Twisted prompts will be applied to contribute parts of the AI-generated essays. AI content pattern analysis and the effect measurement of the twisted prompts will be conducted afterward. To construct models using both LSTM (Hochreiter and Schmidhuber, 1997) and BERT (Bidirectional Encoder Representations from Transformers)(Xian et al., 2024), three experiments will be implemented in this project. Models trained on TF-IDF (Term Frequency-Inverse Document Frequency)("tf–idf," 2024), BERT-based frameworks, and the concatenation of both word embeddings. The main dataset is from the Kaggle competition. The Kaggle dataset contains 10,000 essays with a mix of students' essays and AI-generated essays.

This research addresses five core questions:

1. What is the efficacy of using TF-IDF and BERT for word representation to train LSTM and BERT in distinguishing between AI and human-written texts?
2. How does identifying AI-generated content patterns help with the model training?
3. What's the effect of the "twisted" prompts on the generated text?
4. What is the efficacy of applying twisted prompts to AI-generated essays in distinguishing between AI and human-written texts?
5. What is the efficacy of using the concatenation of TF-IDF and BERT for word representation to train LSTM and BERT in distinguishing between AI and human-written texts?

1. Introduction

Over the past two years, there have been abundant amounts of large language models (LLMs) resources that can generate human-like content, leading to much confusion and disputes around the world (Satpute et al., 2024). In education, educators claim that students' learning performance is largely affected by the convenience LLMs have created (Avetisyan et al., 2023). AI plagiarism refers to the act of using artificial intelligence (AI) tools to generate content—such as text, images, or code—that is then presented as one's own original work without proper attribution to the AI as a source (Garg et al., 2023). These LLMs trained on enormous datasets enable the models to produce content that is very similar to human's performance (Xian et al., 2024). Students take advantage of LLMs to generate essays or contents, leading to deteriorating essential learning milestones (Thompson, 2018). Facing this impediment in education, we developed a model that can tell the difference between the essays produced by middle and high school students and LLMs.

The proposal is organized into key sections: Introduction, background research, dataset and methodology, and project milestones. Initially, the Introduction provides an overview of its motivation and primary objectives. Following this, Section 2 delves into the real-world examples and the driving forces behind the project. Two literature reviews from the Kaggle competition on AI-content detection are conducted. In the third section, the dataset and outline of the potential solutions applicable to the project are included. The fourth section is detailed about the project's work plan and timeline.

2. Background Research

2.1 Motivation

The difficulty in spotting text generated by AI is known by many (Ma et al., 2020). Two snippets of the essays on the topic of life presented below (Grid 1) intend to demonstrate the issue. One was written by a graduate student from the University of Harvard (Shemmassian, 2024), and the other was from ChatGPT.

Task1	Task2
I believe that humans will always have the ability to rise above any situation because life is what you make of it. We don’t know what life is or why we are in this world; all we know, all we feel, is that we must protect it in any way we can. Buddha said it clearly: “Life is suffering.” Life is meant to be challenging, and living requires consistent work and review. By default, life is difficult because we must strive to earn happiness and success. Yet I've realized that life is fickler than I had imagined; it can disappear or change at any time. Several of my family members left this world in one last beating symphony; heart attacks seem to be a trend in my family. They left like birds; laughing one minute and in a better place the next. Steve Jobs inspired me, when in his commencement address to Stanford University in 2005, he said "Your time is limited, so don't waste it living someone else's life. Don't be trapped by dogma--which is living with the results of other people's thinking." I want to make mistakes because that is how I learn; I want to follow the beat of my drum even if it is "out of tune." The	I've always believed that our spirit is unbreakable, and that within us lies the strength to overcome any hurdle that life throws our way. Life, in its essence, is an enigma, a complex puzzle where the pieces are our experiences and emotions. We may not fully grasp the vastness of our existence or the purpose of our journey, but we feel a deep-rooted instinct to cherish and safeguard it with all our might. As the wise Buddha once said, "Life is suffering." It's a path strewn with obstacles, but it is through overcoming these obstacles that we truly live. Life doesn't hand out happiness and success on a silver platter; they are rewards earned through perseverance and hard work. Yet, life has shown me its unpredictability, its capacity to change or vanish in the blink of an eye. I've seen loved ones depart this world swiftly, leaving behind a silence where once there was laughter. Like birds, they flew away, leaving us in one moment and reaching a peaceful beyond in the next. Steve Jobs, in his 2005 Stanford University commencement address, left an indelible mark on me with his words: "Your time is

important thing is to live without regrets, so when my heart ceases to beat, it will make one last happy note and move on. I want to live my life daily. Every day I want to live. Every morning when I wake up, I want to be excited by the gift of a new day. I know I am being idealistic and young, and that my philosophy on life is comparable to a calculus limit; I will never reach it. But I won't give up on it because I can still get infinitely close and that is amazing. Every day is an apology to my humanity; because I am not perfect, I get to try again and again to "get it right." I breathe the peace of eternity, knowing that this stage is temporary; real existence is continuous. The hourglass of life incessantly trickles on and we are powerless to stop it.	limited, so don't waste it living someone else's life." His words resonate with my soul. I yearn to learn from my mistakes and march to the beat of my drum, however erratic it may seem. Living a life free of regrets is my goal, so when my final moment comes, my heart will sing one last joyful note and gracefully move on. I aim to embrace each day as a gift. Every sunrise brings with it the promise of a new beginning, a fresh start. My youthful idealism paints a picture of life as an unattainable calculus limit - always striving, never reaching, yet eternally fascinating in its pursuit. But the beauty lies in the journey, in getting infinitely close to that ideal. Each day is a testament to my imperfection, a chance to strive for better, to reach for perfection in the impermanent theatre of existence. Time marches on, relentless and unforgiving, yet within its passage lies the opportunity for rebirth and renewal.

GRID 1 SNIPPETS OF THE ESSAYS FROM BOTH HUMAN AND AI

By analyzing both examples in detail, some cues can be found to distinguish the difference between the two essays. In Task 1, written by a graduate student from the University of Harvard, personal anecdotes such as "family members experiencing heart attacks" indicate human authorship, providing the personal touch that AI often fails to replicate. Phrases like "Every day is an apology to my humanity" and comparing life's philosophy to "a calculus limit" express deep philosophical feelings in the writing. The way writers show an understanding of abstract concepts in writing remains a challenge for AI. It is also rare to see AI use cultural and famous references in producing content. The student mentioned both Steve Jobs and his speech in the essay to reflect the influence on the author's life. Furthermore, applying metaphors in writing is also evidence to differentiate the two essays. Metaphors like "life as an "hourglass" and existential musings about life and eternity" are difficult for AI to learn and produce. To be more specific, the second essay generated from ChatGPT is produced using lots of similar structure and phrasing. Multiple uses of "life is suffering" and "living a life free of regrets" seem to follow a predictable approach to expressing ideas. AI-generated text often tends to reveal such patterns, using similar repetitive common phrases to convey information. Besides this, the AI content is usually focused on generalization instead of specificity. For instance, it talks about loved ones departing "swiftly," lacking the specific anecdotal feel of Task 1's "family members left this world in one last beating symphony."

ChatGPT is getting better and better at sounding like a human, making it harder for these detectors to keep up (Wang et al., 2019). As ChatGPT's abilities grow, older methods aren't quite cutting it anymore. Therefore, many people are in a race to develop detection tools that can evolve just as fast as ChatGPT and other AI assistant tools. The true challenge lies in ensuring we don't mistakenly identify genuine, human-crafted essays as products of artificial intelligence. The project aims to be fair and accurate because wrongly accusing someone's genuine work of being AI-generated could have serious ethical, legal, and professional consequences (Jiang et al., 2020).

2.2 Literature Review

In this section, two comprehensive literature reviews will be provided. Two teams from the competition, 'LLM - Detect AI-Generated Text', implemented two distinct ways for training the models that performed well in distinguishing between human and AI-generated text.

2.2.1 Team1 - LLMLab

The team proposed a two-pronged approach to AI-generated content detection. The analytical strength of TF-IDF with the contextual understanding capabilities of models like DeBERTa v3 (He et al., 2023) and RoBERTa (Liu et al., 2019). Transformer-based models, such as BERT and its variants leverage the bidirectional encoding method to language understanding and transformer architecture (Vaswani et al., 2023) for processing text. These models are initially pre-trained on a large corpus of text. The pre-training process involves learning to predict masked words (Masked Language Model, MLM) ("Masked language modeling," 2024) and, for BERT specifically, understanding the relationship between sentence pairs (Next Sentence Prediction, NSP) (Sun et al., 2022). After the pre-training stage, the model can be fine-tuned with additional layers based on specific NLP(Natural Language Processing) tasks. ("LLM - Detect AI Generated Text," 2024).

TF-IDF uses the formula below to calculate the weight for each word as the importance of the relationships of words within a document.

Term Frequency (TF):

$$Tf(t,d) = \frac{Number\ of\ times\ term\ t\ appers\ in\ document\ d}{Total\ number\ of\ terms\ in\ document\ d}$$

Inverse Document Frequency (IDF):

$$IDF(t,D) = \log\left(\frac{1Total\ number\ of\ document\ in\ corpus\ D}{Number\ of\ documents\ with\ term\ t\ in\ them}\right)$$

TF-IDF Score:

$$TFIDF(t,d,D) = TF(t,d) \times IDF(t,D)$$

Implementing the TF-IDF pipeline as a reverse-engineering deobfuscation process, the team accurately corrects the systematic spelling errors. This approach serves as a post-processing element that leverages clustering techniques to refine the detection of LLM-generated texts based on similarity metrics.

Traditionally, a TF-IDF pipeline in NLP is produced by converting a text document into a matrix of TF-IDF features, quantifying how important a word is to a document in the corpus whereas, in the reverse engineering approach for error correction, it refers to a methodical process of analyzing text data to detect the pattern or anomalies that show systematic errors and then devising strategies to correct these errors. The reverse engineering process could involve first identifying uncommon variations of commonly used words, then mapping these variations back to their correct form based on the earlier-generated TF-IDF matric, and lastly reapplying the TF-IDF transformation to update the matrix based on the corrected text.

The BERT-based pipeline incorporates two DeBERTa v3 large models and one RoBERTa model, both trained on an overlapping variant of highly diverse datasets to elevate the solution's detection capabilities. The datasets were generated from a large number of open source LLMs of varying sizes between 7-70B parameters, as well as commercial LLM offerings, and a wide variation of temperature and top_p. Generated texts were derived from essay generation, paraphrasing, and text completion prompts. Human texts were curated from student essays and open-source web text. 250k-700k total samples were in each dataset version. The best single BERT-based classifier achieved 0.96 public / 0.915 private leaderboards.

In short, a comprehensive approach to AI-generated content detection is presented by the dual-pipeline which blends the efficiency of TF-IDF features with the depth of understanding afforded by Transformer-based models. Innovative techniques including the curation of datasets, the use of deobfuscation and post-processing techniques, and the strategic training of BERT-based classifiers on diverse texts collectively show the advancement in the field of AI plagiarism detection.

2.2.2 Team2 - Nicholas Broad

Nicholas Broad and his team members have successfully designed an AI-content detection model trained by applying "datamix" techniques. The dataset has been characterized by its size, diversity, and complexity, which enhances the model's generalization capabilities. The team generated AI essays mostly from open-source LLMs such as GPT-3.5 ("ChatGPT," 2020), GPT-4, and synthetic datasets, T5 (Raffel et al., 2023). Human texts were sourced from Persuade corpus (Crossley et al., 2022) and diverse repositories such as OpenAI GPT2 (Radford et al., 2019) output dataset and Wikipedia. ("LLM - Detect AI Generated Text," 2024)

LLM fine-tuning and custom Tokenization were conducted afterward. Fine-tuning with (Q)LoRA (Dettmers et al., 2023) on the Mistral 7b model (Jiang et al., 2023) and DeBERTa-V3 was incorporated. The team also proposed an innovative tactic that combined the custom tokenizer with Masked Language Modeling (MLM) and pseudo-labeling (Lee, 2013), aiming at deepening the model's comprehension of the dataset. Moreover, the team also explored the Ghostbuster approach (Verma et al., 2024) and Ahmet's Unsupervised Approach (Üstün et al., 2021) to experiment with a range of techniques to optimize performance. The team was able to obtain a more precise and accurate prediction for determining whether a text was generated by applying ensemble strategies and transforming raw prediction values into rankings and then taking the average of those rankings across models.

3. Dataset and Methodology

3.1 Dataset

The dataset is from the Kaggle competition, "LLM - Detect AI-Generated Text". Around 10,000 students and AI-generated essays with 7 different academic prompts are included in the dataset. There are two portions of the dataset – training and hidden test set from the other prompts. Most of the training data is from students-written essays, implying the requirement for AI-generated essay generation. Since we don't have access to the hidden test set for the model validation, the testing process will be conducted by the self-generated dataset.

3.2 Methodology

Different LLMs such as ChatGPT-4, and Google Gemini, and open-source models such as LangChain (“langchain 0.1.17, ”2022) will be used to generate AI-produced content for the training data. Each model will be given academic prompts to expand the training data to balance the number of human-written essays and AI-generated essays. Academic prompts will be chosen from the competition’s dataset. The first experiment is expected to be completed to evaluate the LSTM and BERT models generated from TF-IDF and BERT word embedding respectively.

It’s common to see students or users using special prompts to get LLMs to generate more human-like content. To train the model processing the ability to differentiate these contexts, the model will also be trained on the essays generated by these twisted prompts to test the validation of the classifier in the second experiment. The initial experiment used both original and twisted prompts was conducted as shown in Grid 2. In the left essay, the original prompt, "For this essay topic: an explanatory essay to inform fellow citizens about the advantages of limiting car usage. Write me a middle school level essay. " is used. In the right assy, the twisted prompt " For this essay topic: an explanatory essay to inform fellow citizens about the advantages of limiting car usage. Use the same writing style as this. Write it as if it was written by a middle school student. I want it to seem a real human work, not AI-generated content. " is applied.

Explanatory Essay: Advantages of Limiting Car Usage (With the original prompt)	Explanatory Essay: Why We Should Use Cars Less (With the twisted prompt)
In our modern world, cars are everywhere, and they make getting from one place to another easy and comfortable. However, the widespread use of cars has also led to significant problems, such as pollution, traffic jams, and the depletion of natural resources. This essay explores the advantages of limiting car usage, which include environmental benefits, improved public health, and enhanced community life. One of the most important reasons to limit car usage is to protect the environment. Cars emit gases like carbon dioxide and nitrogen oxides, which contribute to global warming and air pollution. These emissions can cause climate change, which leads to extreme weather conditions, rising sea levels, and harmful effects on wildlife. By using cars, we can reduce these emissions and help preserve the planet. Additionally, cars cause noise pollution, which can be reduced by limiting car usage, leading to quieter and more peaceful neighborhoods. Reducing car usage can significantly improve public health. First, decreased emissions lead to cleaner air, which benefits everyone's	Hey everyone! Did you know that reducing how much we use cars can make a huge difference in our community and the world? It sounds like a big claim, but it's true! Here are some awesome reasons why we should think about leaving the car at home more often. Cars contribute a lot to pollution. They release gases that not only make the air dirty but also add to global warming. That's when the Earth gets too warm, causing all sorts of weird weather and harming animals and plants. If we drive less, we help cut down on these gases and keep our planet cooler and cleaner. Also, less car noise means we can actually hear the birds sing and not just horns all the time! Driving less means we have a chance to move more. Instead of just sitting in a car, we could be walking or biking. It's fun, and it helps keep us fit. Clean air also means healthier lungs for us and everyone in our town, especially for kids and old people. So, riding a bike or walking not only makes us stronger but also keeps the air fresh.

health but is especially crucial for children, the elderly, and those with respiratory conditions. Moreover, encouraging alternatives to car travel, such as walking, cycling, or using public transportation, can increase physical activity. This helps in reducing obesity rates and preventing health issues like heart disease, stroke, and certain types of cancer. Limiting car usage can also enhance community life. When fewer cars are on the road, streets become safer, making it easier for people to walk or bike. This can lead to more interactions among community members, strengthening social ties and promoting a sense of belonging. Furthermore, less traffic and easier parking can make local businesses more accessible and attractive, potentially boosting the local economy. To effectively limit car usage, communities should encourage and support alternative forms of transportation. This can be done by improving public transportation systems, making them more reliable, frequent, and affordable. Cities can also invest in safer and more extensive networks of paths for walking and cycling. Additionally, implementing carpooling programs and promoting telecommuting can reduce the number of cars on the road. Limiting car usage has several benefits that can lead to a healthier environment, improved public health, and a stronger community. By understanding and supporting these benefits, we can make informed decisions that lead to a better future for all. Encouraging alternative transportation methods and making small changes in our daily habits can have a big impact on our world. As responsible citizens, it's up to us to take action and make these positive changes happen.	When there are fewer cars on the road, our streets are safer and nicer to be around. We can play more outside, meet our friends on the way to school, and not worry about so many cars rushing by. It makes our neighborhood a friendlier place, where it's easy to stop and chat with someone you know. Plus, it's easier to get to local shops or parks when the streets aren't jam-packed with cars. There's a lot we can do! Maybe we can walk or bike to school if it's not too far. We can ask our parents to think about carpooling—that's when you share rides with friends or neighbors. That way, fewer cars are on the road. Also, we can try to use buses or trains more. They're there for a reason, right? So, driving less is about more than just saving gas. It's about making our air cleaner, keeping ourselves healthier, and making our neighborhood a better place to live. It might seem like a small change, but if all of us try to do a little bit, it can make a big difference. Let's try it out and see how much better we can make things by just choosing not to use the car sometimes. Thanks for thinking about this with me! Let's make some changes together!

GRID 2: THE INITIAL EXPERIMENT ON TWISTED PROMPTS

The result shows obvious differences between the original prompt and the twisted prompt. Without the twisted prompt, the tone in the left essay is formal and structured whereas a more casual and conversational tone is used in the right essay with the twisted prompt. Also, technical terms and academic language such as "depleted natural resources" and "nitrogen oxides" are produced without the twisted prompt, allowing the essay to seem more academic and professional. In the right essay, engagement feeling and a personal touch writing styles are integrated. Sentences such as "Did you know?" or "Thanks for thinking about this with me! " elevate the overall interactive feeling towards the audience. Another observation is that the suggestions provided in the left essay often involve community-wide or policy-level changes, such as improving public transportation systems and promoting telecommuting. However, in the right essay, suggestions are more personal and focus more on an individual or family level, such as walking or biking to school, which might be more relatable and feasible for middle school students.

We are interested in both the effect of the "twisted" prompts on the generated text and the AI content patterns found from the analysis. The main task of finding these variations is to improve AI's capability to produce text similar to human-written texts and, as a result, increase performance for the model improvement on detecting AI-generated text. The effect of the "twisted" prompt will be evaluated on the length of the word, the length of the sentence, the use of personal pronouns (e.g., I or you), spelling mistakes, the use of other parts of speech, grammatical constructions (e.g., passive or active voice), or punctuation usage (Figure 1). AI patterns such as the presence of human authorship, philosophical sentiment, abstract concepts, cultural and famous references, and metaphors will be used to develop a multi-dimensional scoring function (Figure 2). The purpose of this is to see whether these patterns are associated with the label in the labeled data to further improve the classification results.

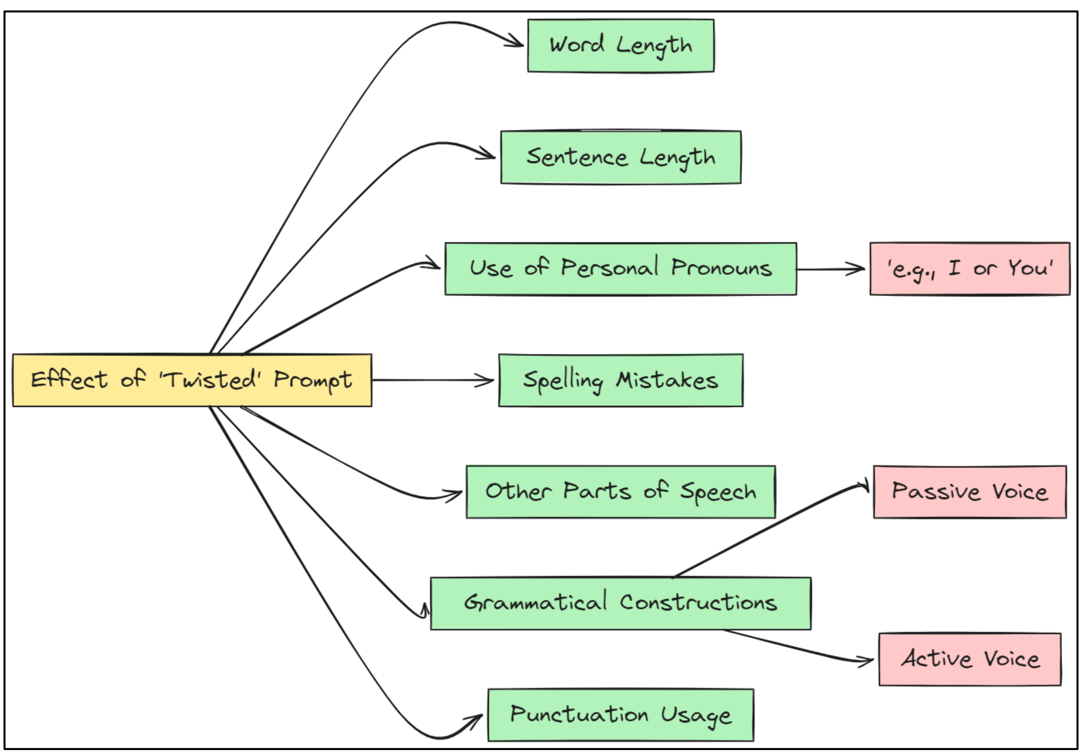

FIGURE 1 : HOW TO EVALUATE THE EFFECT OF THE "TWISTED" PROMPT

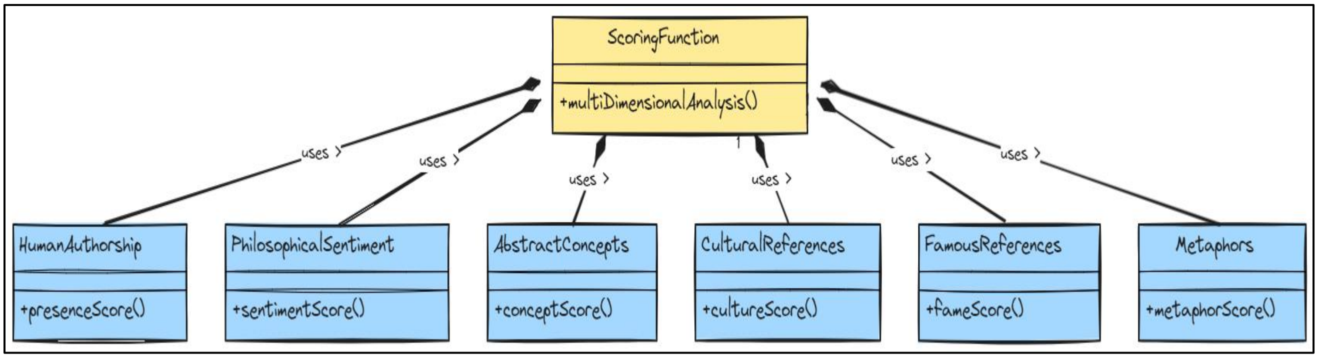

FIGURE 2: MULTI-DIMENSIONAL SCORING FUNCTION FOR AI PATTERNS

The fine-tuning process is applied in the second experiment to improve the model's generalization for the input data. The third experiment will be focused on the concatenated TF-IDF and BERT word embedding to train the models. The goal is to see how better performance both models will yield after the process. Some other neural or pre-trained large language models will be explored at this stage. The overview process of the project is provided in Figure 3.

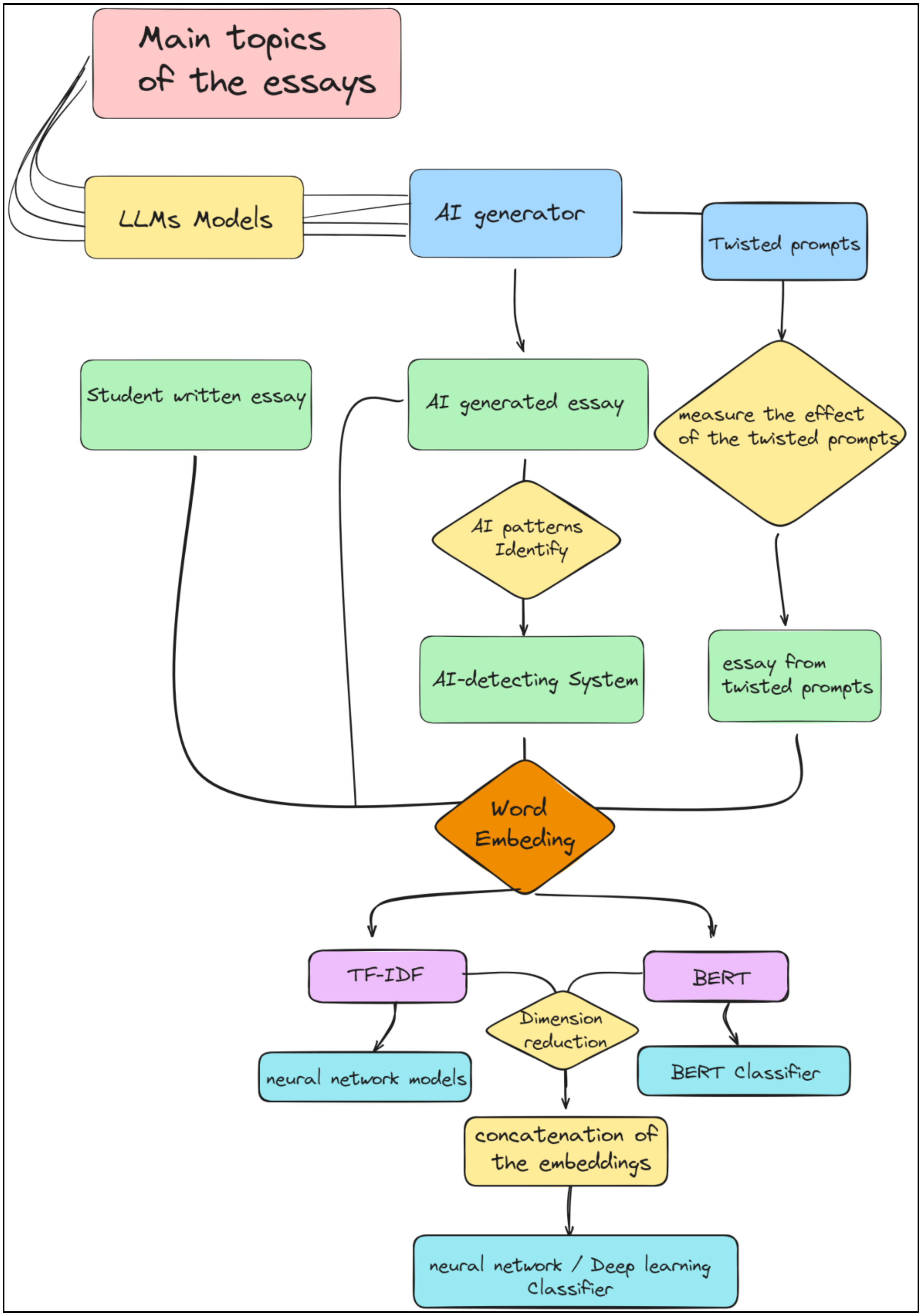

FIGURE 3: THE OVERVIEW PROCESS OF THE PROJECT

4. Project Milestones

This section outlines the schedule for completing this project. The table presented below (Grid 3) details the timeline and the specific tasks planned for this project.

Date	Tasks
17/5 – 30/5	● Research different methods for the AI-content detection classifier.
30/5 – 13/6	● Based on the prompts, produce AI-generated content using LLMs. ● Analyze the AI content patterns ● Run the first experiment on TF-IDF and BERT word representations and feed to the LSTM and BERT model. ● Meeting with the supervisor on 13/6
13/6 – 20/6	● Embed parts of the twisted content from LLM to increase the diversity of the training data. ● Evaluate the effect of the "twisted" prompts on the generated text ● Run the second experiment on both LSTM and BERT models on the new training data. ● Fine-tuning the models to improve the performance. ● Meeting with the supervisor on 20/6
20/6 – 27/6	● Run the third experiment on the new concatenated training data from both TF-IDF and BERT. ● Explore some other models besides LSTM and BERT for potential improvement. ● Meeting with the supervisor on 27/6
27/6 – 4/7	● Finish the Introduction, literature review, and methodology of the dissertation. ● Meeting with the supervisor on 4/7
4/7 – 11/7	● Finish the Data Analysis, Results and Discussion, and Conclusion of the dissertation. ● Meeting with the supervisor on 11/7
11/7 – 31/7	● Dissertation revision ● Presentation preparation ● Meeting with the supervisor on 25/7

GRID 3: THE SCHEDULE FOR THE PROJECT

From 17/5 to 30/5, two weeks will be spent researching different methods for AI-content detection classifiers. Literature reviews are planned to be completed at this stage for future analysis and model construction. From 30/5 to 13/6, AI-generated essays will be produced from LLMs to balance the size of our training data. The first experiment will be conducted on TF-IDF and BERT word representation for the models LSTM and BERT. I will have the first meeting with Professor Julie Weeds to discuss the past process and further actions. From 13/6 to 20/6, apart from the training data, twisted prompts will be embedded from various LLM models to increase the diversity of the training data. We will then use the new training data for the second experiment on both LSTM and BERT models. Fine-tuning processes are also planned to be conducted to improve the performance. The second meeting with the supervisor is scheduled around 20/6. From 20/6 to 27/6, I will run the third experiment on the new concatenated training data from both TF-IDF and BERT word representation. Exploration for other models besides LSTM and BERT for potential improvement will be also done at this point. The third meeting with the supervisor will be around 27/6. From 27/6 to 4/7, the focus will be placed on finishing parts of the dissertation including introduction, literature review, and methodology. The fourth meeting with the supervisor is around 4/7. After 4/7, we will continue writing the dissertation. Parts planned to be completed are data analysis, results and discussion, and conclusion. The fifth meeting with the supervisor will be right after all parts have been completed. For the following weeks, I will focus on revising the content of the dissertation and preparing for the preparation. The sixth meeting will be scheduled around 31/7.

Bibliography

Avetisyan, K., Malajyan, A., Ghukasyan, T., Avetisyan, A., 2023. A Simple and Effective Method of Cross-Lingual Plagiarism Detection.

ChatGPT [WWW Document], n.d. URL https://chat.openai.com (accessed 4.10.24).

Claude [WWW Document], n.d. URL https://claude.ai/login?returnTo=%2F%3F (accessed 5.3.24).

Crossley, S.A., Baffour, P., Tian, Y., Picou, A., Benner, M., Boser, U., 2022. The persuasive essays for rating, selecting, and understanding argumentative and discourse elements (PERSUADE) corpus 1.0. Assess. Writ. 54, 100667. https://doi.org/10.1016/j.asw.2022.100667

Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L., 2023. QLoRA: Efficient Finetuning of Quantized LLMs. https://doi.org/10.48550/arXiv.2305.14314

Garg, H., Sain, J., Rawat, A., 2023. AI PLAGIARISM: A COMPREHENSIVE ANALYSIS OF AUTOMATED CONTENT 11, 99–105.

He, P., Gao, J., Chen, W., 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. https://doi.org/10.48550/arXiv.2111.09543

Hochreiter, S., Schmidhuber, J., 1997. Long Short-Term Memory. Neural Comput. 9, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D. de las, Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.-A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E., 2023. Mistral 7B. https://doi.org/10.48550/arXiv.2310.06825

Jiang, X., Xu, S., Zhang, S., Cao, S., 2020. Arbitrary-Shaped Text Detection withAdaptive Text Region Representation. IEEE Access 8, 102106–102118. https://doi.org/10.1109/ACCESS.2020.2999069

langchain 0.1.17 — 🦜🔗 LangChain 0.1.17 [WWW Document], n.d. URL https://api.python.langchain.com/en/latest/langchain_api_reference.html (accessed 5.3.24).

Lee, D.-H., 2013. Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks. ICML 2013 Workshop Chall. Represent. Learn. WREPL.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. https://doi.org/10.48550/arXiv.1907.11692

LLM - Detect AI Generated Text [WWW Document], n.d. URL https://kaggle.com/competitions/llm-detect-ai-generated-text (accessed 4.8.24).

Ma, C., Sun, L., Zhong, Z., Huo, Q., 2020. ReLaText: Exploiting Visual Relationships for Arbitrary-Shaped Scene Text Detection with Graph Convolutional Networks.

Masked language modeling [WWW Document], n.d. URL https://huggingface.co/docs/transformers/en/tasks/masked_language_modeling (accessed 5.5.24).

Nguyen, Q.H., 2023. AI and Plagiarism: Opinion from Teachers, Administrators and Policymakers. Proc. AsiaCALL Int. Conf. 4, 75–85. https://doi.org/10.54855/paic.2346

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., n.d. Language Models are Unsupervised Multitask Learners.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J., 2023. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.

Shemmassian, D., 2024. 14 College Essay Examples From Top-25 Universities (2024–2025) [WWW Document]. Shemmassian Acad. Consult. URL

https://www.shemmassianconsulting.com/blog/college-essay-examples (accessed 4.10.24).
Sun, Y., Zheng, Y., Hao, C., Qiu, H., 2022. NSP-BERT: A Prompt-based Few-Shot Learner through an Original Pre-training Task —— Next Sentence Prediction, in: Calzolari, N., Huang, C.-R., Kim, H., Pustejovsky, J., Wanner, L., Choi, K.-S., Ryu, P.-M., Chen, H.-H., Donatelli, L., Ji, H., Kurohashi, S., Paggio, P., Xue, N., Kim, S., Hahm, Y., He, Z., Lee, T.K., Santus, E., Bond, F., Na, S.-H. (Eds.), Proceedings of the 29th International Conference on Computational Linguistics. Presented at the COLING 2022, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, pp. 3233–3250.
tf–idf, 2024. . Wikipedia.
Thompson, V., 2018. Detecting Cross-Lingual Plagiarism Using Simulated Word Embeddings. https://doi.org/10.48550/arXiv.1712.10190
Üstün, A., Berard, A., Besacier, L., Gallé, M., 2021. Multilingual Unsupervised Neural Machine Translation with Denoising Adapters, in: Moens, M.-F., Huang, X., Specia, L., Yih, S.W. (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp. 6650–6662. https://doi.org/10.18653/v1/2021.emnlp-main.533
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2023. Attention Is All You Need. https://doi.org/10.48550/arXiv.1706.03762
Verma, V., Fleisig, E., Tomlin, N., Klein, D., 2024. Ghostbuster: Detecting Text Ghostwritten by Large Language Models. https://doi.org/10.48550/arXiv.2305.15047
Wang, X., Jiang, Y., Luo, Z., Liu, C.-L., Choi, H., Kim, S., 2019. Arbitrary Shape Scene Text Detection with Adaptive Text Region Representation.
Xian, J., Yuan, J., Zheng, P., Chen, D., 2024. BERT-Enhanced Retrieval Tool for Homework Plagiarism Detection System.