



Department of Computer Science
BSc (Hons) Computer Science
(Artificial Intelligence)

Academic Year: 2022/2023

Social Media User Impersonation:
Evaluating the Advancements of
Generative Models

Karlís Kanepe, 2022092

A report submitted in partial fulfilment of the requirements for the degree of
Bachelor of Science

Brunel University London
Department of Computer Science
Uxbridge
Middlesex
UB8 3PH
United Kingdom
T: +44 1895 203397
F: +44 (0) 1895 251686

I certify that the work presented in the dissertation is my own unless referenced.

Signature: Karlis Kanepe

Date: 06.04.2023

Total Words: 7692

Abstract

The rapid progression of generative models, combined with the increasing amount of public data, has led to many innovative content generation techniques. This necessitates a critical examination of their potential applications for malicious misuse, particularly in relation to social media user impersonation. The effectiveness of these models in mimicking social media users remains largely unexplored.

This research aims to evaluate the capabilities of generative models in social media impersonation by employing advanced algorithms to predict user responses and generate new posts using publicly accessible Twitter data. The study will evaluate the model's capacity to generate authentic content using both quantitative and qualitative indicators. Findings from this study will provide insights into the strengths and limitations of generative models in the context of social media impersonation and illuminate opportunities for future research and development in this rapidly evolving domain.

Acknowledgements

I would like to convey my deepest appreciation to my tutor, Professor Derek Groën, whose unwavering support, advice, and useful insights have been invaluable in my academic path. His sincere passion for education and commitment to his pupils have made him a remarkable mentor.

I am also thankful to Dr. Alina Miron, who, as an examiner of my work, gave insightful feedback that significantly shaped the trajectory of my project. Her knowledge of deep learning has been invaluable in shaping my grasp of the topic.

Furthermore, I want to express my sincere gratitude to the open-source community, including developers and supporters, who have freely given the tools and resources that enabled innovation worldwide without capitalising on it. Their dedication to making knowledge available and promoting cooperation has been genuinely amazing, and it has played a critical role in the project's success.

Table of Contents

Abstract	i
Acknowledgements	ii
List of Figures.....	v
1. Introduction.....	1
1.1 Aims and Objectives	1
1.1.1 Aims	1
1.1.2 Objectives.....	2
1.2 Approach	2
1.3 Project Outline	3
2. Background.....	5
2.1 Deceiving Machines	5
2.1.1 Role of Imitation	5
2.1.2 Approach to Deception	5
2.1.3 Not Always Malicious	6
2.2 Generative Models	6
2.2.1 Small Language Models.....	6
2.2.2 Large Language Models.....	7
2.3 Success Criteria	7
2.3.1 BLEU Score	7
2.3.2 Language Components.....	8
2.4 Impersonation research.....	8
2.4.1 Finding Impostors.....	8
2.4.2 Impostor GAN	9
3. Methodology	10
3.1 Waterfall Methodology	10
3.2 Agile elements	10
3.3 Preparation	11
3.4 Model training	12
3.5 Measurements and insights.....	12
3.6 Result communication	13
4. Design	14
4.1 Data Collection.....	14
4.2 Generative Model Design	14
4.3 Model Testing and Evaluation Design.....	16

4.4 Technology Selection	16
5. Implementation	18
6. Evaluation	19
6.1 Quantitative Analysis	20
6.1.1 Size Comparison	20
6.1.2 Perplexity	20
6.1.3 BLEU Score	21
6.1.4 Cross Entropy Loss	22
6.2 Qualitative Analysis.....	25
6.3 Comparative Analysis.....	26
6.3.1 Model Evaluation	26
6.3.2 Strengths and Limitations	26
7. Conclusion	28
7.1 Summary of Findings	28
7.2 Future Work.....	28
References.....	30
Appendix A Personal Reflection	32
A.1 Reflection on Project	32
A.2 Personal Reflection.....	32
Appendix B Ethics Documentation	34
B.1 Ethics Confirmation	34

List of Figures

Figure 1. A Gantt chart of the Final Year Project	10
Figure 2. Agile Elements.....	11
Figure 3. A flowchart depicting the training process.	15
Figure 4. A flowchart outlining the process of testing.	15
Figure 5. Complexity and problems.....	18
Figure 6. Generated content examples.	19
Figure 7. Small language model parameter comparison.....	20
Figure 8. Line chart of perplexity per model.	21
Figure 9. BLEU score per model.....	22
Figure 10. RNN loss function.	23
Figure 11. GRU loss function.	24
Figure 12. GPT-2 loss function.....	24
Figure 13. Human evaluation average by model.....	26

1. Introduction

Deception lies at the heart of artificial intelligence; simple machines do not possess consciousness (Dehaene, Lau & Kouider, 2021), nevertheless the ultimate goal for human engineers is to create a system that can successfully deceive another human. The Turing test is an early example of this goal, as AI was charged with convincing a human that they were conversing with another human rather than a machine. This paper delves into the world of deception, concentrating particularly on impersonation—the ability to learn, imitate, and exploit specific behaviours to deceive others.

In today's digital age, an individual's data can often be publicly available, but many people are unbothered about the consequences of such transparency. Data is power, and with power comes the ability to manipulate and control, which is why large social media companies can offer their platforms for free; in this context, the user becomes the product. Some argue that AI has already surpassed the Turing test, citing examples such as the engineer testing Google's AI system and trying to alert the public to its perceived consciousness (Will Oremus, 2022). The potential ramifications of malicious organisations exploiting this technology for political and economic benefit without regard for ethical ramifications are frightening.

The objective of this study is to look at the capabilities of deep learning architectures in mimicking social media user behaviour in order to identify key differences in performance, credibility, and logic. With the goal that by investigating these issues, we would be able to raise awareness about the hazards connected with generative models and the possibility for impersonation in cyberspace, using data that is openly available to everyone on the world.

1.1 Aims and Objectives

1.1.1 Aims

The primary purpose of this project is to educate social media users about the possible threats they face, such as identity theft and tailored ad targeting, by exhibiting the capabilities of various generative models in impersonating individuals on limited, publicly available data. Furthermore, this study aims to assess the efficacy of multiple language model designs and sizes to accomplish the goal of impersonation, highlighting the specific strengths and limitations of each model to give a comprehensive understanding of their performance.

1.1.2 Objectives

- Conduct a thorough literature review to understand the fundamentals of impersonation, investigate deception, and identify appropriate architectures and techniques for the project.
- Collect and curate a dataset from Twitter, ensuring sufficient amount of data and unique user profiles, clean data for optimal usage.
- Develop and implement different generative models using a variety of architectures in addition to fine-tuned pre-existing models. Optimise the models to improve accuracy and related performance metrics by adjusting parameters and configurations.
- Create a clear set of success criteria that are aligned with the research objectives, serving as a benchmark for assessing the performance of each model.
- Conduct quantitative and qualitative analyses of the model outputs, comparing their efficacy and emphasising key insights within the context of the aim.
- Clearly communicate the findings, emphasising the potential consequences of ignorance on social media platforms and raise awareness about the dangers of online impersonation.

1.2 Approach

1. Administer a thorough literature study that is divided into sections as follows:
 - Examine existing studies on impersonation by generative AI models.
 - Consider user uniqueness, post count, and data collection techniques when looking for suitable input data for training.
 - Investigate various architecture designs and their applicability for this project, then select the most optimal and promising ones for implementation.
2. Snsrape, an open-source tool with customisable options for scraping data from Twitter, is used to collect data.
3. Remove unnecessary material, duplicates, null numbers, and hyperlinks from the data. Partition the data and conduct tokenisation and other pre-processing steps as needed.
4. Create models based on the research:
 - Models built on data-driven architectures, such as RNN, GRU, LSTM, and Transformers, should be created.
 - Fine-tune pre-existing models that are beyond the hardware limitations of this project.

-
- Experiment with alternative approaches, for example, different sizes of LLM, different pre-trained transformer models.
5. Conduct qualitative research by evaluating the performance of the models using human recognition tests. Use measures such as perplexity, accuracy, and consistency to conduct quantitative analysis.
 6. Perform a comparative analysis to highlight important findings. Create clear and appealing data visualisations that demonstrate the performance of each model based on various metrics, highlighting growth rates and advancements.
 7. Present a comprehensive conclusion that summarises findings and makes predictions about future developments to effectively convey the results. Encourage readers to be mindful of the presence of such innovations and how they might impact their privacy.

1.3 Project Outline

Chapter 1. Introduction.

Describes the significance of this field of research, goals, and path leading up to this. Discusses the use of generative models for imitation online and the potential harm that might arise.

Chapter 2. Background.

Includes a thorough discussion of impersonation's ethical implications and a literature review covering generative models, impersonation techniques, and associated architectures.

Chapter 3. Methodology.

Describes how the study was conducted from start to finish, detailing everything from data collection and pre-processing to model creation and evaluation.

Chapter 4. Design.

Explanations of the thought processes behind the model and parameter selections are provided, as well as specifics on the design and configuration of the different generative models.

Chapter 5. Implementation.

Explains how models are created and trained, how problems were solved during execution, and what it's like to put the selected architectures into practise.

Chapter 6. Evaluation.

Results from both qualitative and quantitative assessments are presented, and the performance of each model is discussed in regard to the success criteria and the study's aims and objectives.

Chapter 7. Conclusion.

Provides an overview of the evaluation's key results and a glimpse into the potential future of harmful use of generative models, stressing the need for users to exercise caution while interacting with social media.

2. Background

This background chapter ventures into the complex field of generative models, with a special emphasis on its usage in social media user impersonation. We will investigate key concepts, strategies, and performance metrics related with these models through a comprehensive literature analysis, emphasising their capacity to predict user responses and produce authentic content. The paper will examine existing research on impersonation using different deep learning architectures in order to identify the most optimal and viable ones for implementation. This chapter seeks to lay a solid basis for the research topic, allowing the reader to acquire a better grasp of the study's background and direction.

2.1 Deceiving Machines

2.1.1 Role of Imitation

In his journal article Alan Turing (Turing, 1950) questions artificial intelligence from multiple perspectives. One of these perspectives is his mathematical objections, in which he argues that, like machines, the human brain has limitations. As a result, with time and technological advancements, it would be possible to reach similar levels of complexity in machines as those found in human brains.

Turing also believed that humans could be deceived by machines, even with the technology available during his time. However, he acknowledged that language models of his era were not yet computationally capable of convincingly impersonating adult humans. Instead, he suggested that these models should be judged on their ability to trick people into believing they were interacting with children. This notion is particularly alarming in the context of the present study, as it raises concerns about the potential for large language models to deceive children into engaging in conversations with what they believe to be their peers.

2.1.2 Approach to Deception

In the field of artificial intelligence, two main paradigms have emerged: the model-driven paradigm and the data-driven paradigm (Sarkadi, 2018). The model-driven paradigm involves building AI systems that reason using models containing beliefs and knowledge about the world and other agents, while the data-driven paradigm focuses on building AI systems that reason based on available evidence without the use of such models.

This thesis explores the model-driven approach in an attempt to gain key insights from limited data. As mentioned in the referenced paper, a mental model or a theory of mind is crucial for understanding deception in AI systems. However, incorporating such elements is challenging, and typically requires access to large amounts of data – which this study will be limiting. The paper also reveals that most research in this area has predominantly focused on the data-driven approach. By investigating the model-driven paradigm, this thesis aims to provide a fresh perspective on understanding and addressing deception in AI systems.

2.1.3 Not Always Malicious

As stated by (Rovatsos, 2019) Castelfranchi's work acknowledges that deception can sometimes yield positive outcomes for the deceived, as self-interested agents might employ deception in the pursuit of altruistic goals. This perspective highlights the potential for individuals to use deceptive machines for the greater good of society. When engaging with AI systems like Alexa's and Siri's, the thought of a machine deliberately choosing to deceive a human might make us uneasy. However, there are instances where such deception could bring about positive results.

For example, a deceptive AI system could impersonate a friendly entity, providing companionship or support to someone in need. Additionally, it could imitate a deceased person using publicly available data, offering closure or happiness to grieving individuals. While deceptive machines can undoubtedly pose challenges, it is crucial to consider their potential for positive impact and explore ways to harness their capabilities for the betterment of society.

2.2 Generative Models

2.2.1 Small Language Models

In this study (Mangal, Joshi & Modak, 2019), the authors successfully generated scripts using three distinct deep learning models, namely Bidirectional RNN, LSTM, and GRU. By training these models on a dataset consisting of a famous TV series' dialogues and scene descriptions. A comparative analysis of the models revealed that LSTM was the most efficient in terms of text generation, followed by GRU and then Bidirectional RNN. Conversely, Bidirectional RNN exhibited the lowest loss, followed by LSTM, with GRU exhibiting the highest loss. In terms of text generation time, LSTM was the fastest, with GRU slightly slower and Bidirectional RNN being the slowest.

The research approach taken by the authors resonates with the methodology proposed for this thesis, which also involves a trial-and-error exploration of

parameters to identify the most effective outcomes. It is worth noting that the input data in their study consists of longer texts, which typically require larger models to maintain coherence. As part of this thesis, the same architectures used here will be tested and compared, offering valuable insights into their performance and potential for text generation tasks. The authors suggest that future research could further enhance and optimize these models with more data and computational resources, a direction that in this thesis will not be pursued, however the approach is going to be similar.

2.2.2 Large Language Models

As evidenced in the research (Montesinos, 2020), transformer-based models such as BERT and GPT-2, when utilized with larger datasets, have displayed remarkable superiority in comparison to RNNs and other similar models. These models, built upon advances in attention mechanism (Vaswani et al., 2017), have achieved state-of-the-art performance on a wide range of language modelling benchmarks, this success is accredited to parallelization as they aren't sequential, but RNN models are. These advances also allow transformers models to be trained a lot faster.

The development of such models has continued to accelerate. Even though this research is not recent, the remarkable achievements of these models back then serve as an indication of the potential for even more ground-breaking advancements in the field of large language models (LLMs). As the field continues to evolve, we can expect to see further improvements in these models and their applications, as they increasingly become an integral part of natural language processing and understanding.

2.3 Success Criteria

2.3.1 BLEU Score

The development of the BLEU metric (Papineni et al., 2001) has significantly accelerated the evaluation process in the field of natural language processing and machine translation. Despite being a relatively older invention, it continues to be a relevant and widely used tool for assessing the performance of NLP models. The BLEU metric employs a sophisticated formula that enables it to evaluate text more effectively than human evaluators, making it a valuable resource for researchers and developers working on language models.

This innovative metric serves as a quantitative analysis tool, providing a comprehensive and objective measure of translation quality. By considering various

aspects of the translated text, such as n-gram matches and modified precision, the BLEU score offers a reliable and language-independent means of comparing machine-generated translations to human-generated reference translations. Its efficiency and accuracy have made it an indispensable tool in the ongoing development of increasingly advanced natural language processing models.

2.3.2 Language Components

In the realm of cognitive research, the intention behind a text can be challenging to measure, given the vast array of methods available for classifying and evaluating various aspects of cognition. In response to this challenge, the paper by (Alvarado et al., 2002) proposes several approaches that aim to capture distinct dimensions of cognitive processing. By incorporating these diverse methodologies, a more comprehensive understanding of language can be achieved.

The first approach focuses on tests of mental representation, which are crucial in understanding how individuals form internal representations of their experiences. These tests delve into the nature of classical conditioning effects and explore how expectations are formed, based on the principle that certain behavioural outcomes can only be achieved if specific mental processes are taking place.

Another key aspect of cognitive function involves the representation of the self. In order to assess all context around a person, various tasks and paradigms are employed to evaluate how individuals form self-schemas, which are essentially mental representations of themselves. These self-schemas are shaped by an individual's experiences and can influence their behaviour and decision-making processes.

Lastly, the paper addresses social cognition, which involves understanding how individuals perceive, interpret, and respond to social interactions. To evaluate this domain, tests are designed to assess social encoding processes, social inference, causal attribution, empathy, and attachment. By exploring these various aspects of social cognition, a deeper insight into how individuals navigate and interact within their social environments can be obtained.

2.4 Impersonation research

2.4.1 Finding Impostors

The existing literature on the topic of impersonation primarily revolves around classification papers, with only a handful of studies delving into the creation of

content-generating models. This could be attributed to the ethical implications associated with generating fake content that mimics the behaviour of real users.

This project (Zarei et al., 2020), however, takes a different approach, closely relating to the study mentioned, which focused on scraping publicly available data from social media platforms. The researcher in that case implemented a sophisticated algorithm for detection, utilizing a large language model, BERT, to output vectorized text. This text was then fed into a Convolutional Neural Network (CNN) and connected to an LSTM layer. The final output from this processing determined the type of post, classifying it as 'bot-generated', 'fan-generated', or 'genuine' content.

Drawing inspiration from this approach, the current project aims to further explore content generation and detection within the realm of social media impersonation, expanding upon the techniques and methods employed in the existing body of research.

2.4.2 Impostor GAN

The study (Nafees et al., 2021) in focus adopted an approach inspired by previous models designed for detecting fake accounts, utilizing two generating models and one detector. The researchers expressed their intent to incorporate GPT-3 into future models, a step that is currently being implemented in this project.

Interestingly, the study chose to use Twitter data, specifically targeting political personas. This choice, however, could potentially raise concerns within the community due to the sensitive nature of political content. In terms of evaluation, the study's scope was rather limited, with the primary focus being on the analysis of their training data. The researchers concentrated mostly on refining their approach to achieve improved results, rather than conducting a comprehensive evaluation of the model's performance.

In contrast to the aforementioned study, this project aims to build upon the existing foundation by implementing cutting-edge language models like GPT-3 and expanding the scope of the evaluation process to gain deeper insights into the model's performance and real-world applicability.

3. Methodology

3.1 Waterfall Methodology

The Waterfall methodology is a widely recognized project management approach that emphasizes a linear, sequential structure. This methodology is characterized by its key principles, which involve breaking down the project into distinct phases and completing them in a specific order, ensuring that each phase is completed before moving on to the next. The rationale behind choosing the Waterfall methodology for this project is its ability to provide predictable results within the constraints of a predetermined time frame. This aligns well with the project's need for a structured approach and enables efficient documentation, which is crucial for tracking progress, observe an example of such structure in Figure 1. Furthermore, the Waterfall methodology supports the alignment of the project's objectives by establishing clear goals and requirements that need to be fulfilled in each phase, thereby ensuring that the project remains focused and on track to achieve its desired outcomes.

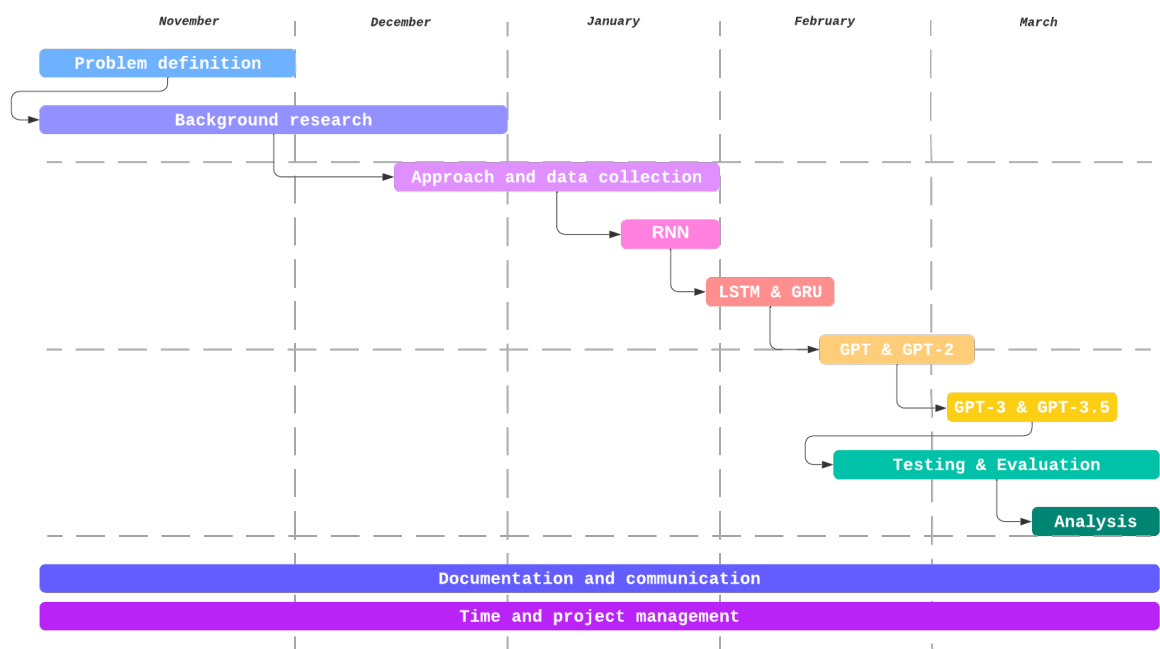


Figure 1. A Gantt chart of the Final Year Project

3.2 Agile elements

There are incorporated some principles from Agile methodologies, specifically the iterative and flexible nature of these approaches. Agile methodologies, emphasize continuous improvement through iterative cycles, enabling teams to adapt and respond effectively to changing requirements and project constraints.

The project's adoption of Agile elements is evident in the development of different generative models and fine-tuning of various pre-trained models. Each model follows a simplified Agile structure, consisting of three stages: Research, Design, and Implementation. In the Research stage the model's architecture and parameters are investigated, ensuring a solid foundation for development. The Design phase involves planning and strategizing the implementation of the model, while the Implementation stage encompasses the actual development and deployment of the model.

By incorporating Agile elements, as shown in Figure 2, into the methodology, the project can rapidly adapt to new information, requirements, and potential challenges. This approach allows for the efficient development of multiple generative models, ensuring that the most effective solutions are identified and applied in the context of the study's objectives.

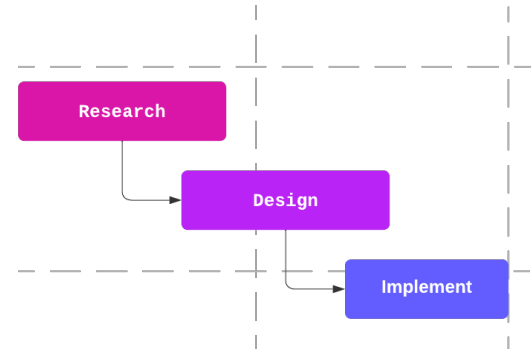


Figure 2. Agile Elements

3.3 Preparation

Projects first step is the preparation for training, which involves a comprehensive literature review and data collection process. The literature review entails an in-depth analysis of existing research to inform the project's direction and decisions. This includes examining the architecture of various generative models, understanding their key differences and parameters, and identifying potential drawbacks and risks associated with each model.

Following the literature review, the data collection process is outlined, focusing on the design and implementation of a suitable method for gathering large quantities of data from Twitter. Several Twitter scraping tools were tested, ultimately selecting the open-source project snsrape, specifically its Twitter module, due to its superior performance with large data sets. The Python scriptable nature of snsrape made it convenient to use for this project. The data collected included both the user's reply to a post and the original post being replied to. In cases where the collected post was an original post, a generic description was hand-labelled to facilitate analysis. This approach ensures that the project is grounded in a solid understanding of the existing research and employs an effective data collection method to achieve its goals.

3.4 Model training

Several critical procedures are done here to guarantee that the generative models are developed and implemented effectively. To begin, data pre-processing is carried out using Python libraries such as Pandas and Regex to clean and optimise the acquired data for later use in model training.

Following that, Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) architectures are the first to be put on trial. These models are taught using a personal MacBook, which has the processing capability to carry out the training procedure in a reasonable amount of time.

Following model development, optimisation is performed by retraining or restarting the models with changing parameters such as layer count, learning rate, and hidden layer count. The goal of this approach is to refine the model's, in order to increase their performance.

Finally, pre-trained models are fine-tuned by accessing bigger models via APIs, such as HuggingFace's API for the GPT-2 model, following a tutorial posted by (Pointer, 2020). Continuing by closely following the documentation of (OpenAI, 2023) for fine-tuning the GPT-3 model. These models are fine-tuned using the exact same dataset as the previously produced RNN and GRU models, ensuring a consistent foundation for comparison and analysis, although it is converted to a different format for training it for GPT-3. Overall, these methodological procedures establish a solid foundation for the construction, optimization, and assessment of diverse generative models within the scope of the project.

3.5 Measurements and insights

Following a comprehensive approach to evaluate the generative models, using both quantitative and qualitative methods to assess their performance. For the quantitative evaluation, a set of prompts is utilized to gather results, which are then analysed in terms of perplexity, validity, and BLEU score, providing an objective measure of the models' effectiveness. In addition to these quantitative metrics, a simple survey is conducted by the researcher to qualitatively evaluate the models, capturing subjective insights and impressions on the performance of each model.

Conducting an individual evaluation of each model, a comparative analysis is performed to draw meaningful insights and identify trends. By combining the scores from the various tests, averages for each model are calculated and visually

represented in graphs, facilitating the examination of relationships, correlations, and patterns among the models. Through this comprehensive evaluation process, the study can effectively compare the strengths and weaknesses of different generative models, thereby guiding the selection of the most suitable solution for the research objectives.

3.6 Result communication

The interpretation of the data, the drawing of conclusions, and the contemplation of the larger implications within the framework of the study goal are the primary focuses at this point. In this part, we go deeper into the significance of the findings and investigate the ways in which they contribute to our knowledge of generative models and the possible influence these models might have on users of social media. The study gives insights about the current status of AI impersonation as well as its prospective consequences. These insights are provided by examining the outcomes and comparing them to the larger context.

4. Design

4.1 Data Collection

The data collection aspect of the project involved a carefully planned approach, beginning with the selection of Twitter as the data source. Twitter was chosen due to its short-form text-based data, which is ideal for chat-like prompting and has been popularized by models like ChatGPT. Moreover, the platform's predominantly public nature allows for easier access to user data.

The data collection tool utilized in this study is Snsrape, which offers significant advantages in terms of customizability and efficiency. Notably, Snsrape does not rely on Twitter's API, which often imposes bandwidth restrictions and slows down data collection. Instead, it offers a more streamlined and faster method for gathering data.

For the sampling strategy, the project aimed to select unique users and gather a sufficient amount of data for training the models. Users were chosen based on their posting frequency (at least weekly) and prominence on the platform. High-profile individuals such as Barack Obama, Elon Musk, Bill Gates, and Jack Dorsey were included, as their large follower bases necessitate maintaining high-quality content, resulting in less data cleaning required. However, due to time and computing power constraints, the models were ultimately trained solely on Elon Musk's data, as his posts exhibit distinct patterns that proved valuable for the study.

Cleaning and pre-processing were also employed to ensure optimal performance. Depending on the model, the process was different, but the basics were similar - First, the Python library Pandas was employed to read and concatenate the data, organizing it into a structured format. Next, data cleaning was conducted using the Python library Regex, which enabled the removal of unnecessary characters, inconsistencies, and irrelevant information. Finally, pre-processing involved extracting unique characters from the dataset and converting them into PyTorch tensors using the PyTorch library. This transformation allowed for efficient and seamless integration with the generative models during the training phase, ultimately contributing to the project's success.

4.2 Generative Model Design

The model design was a critical aspect of this project, requiring careful selection of architectures and configurations. The decision-making process involved choosing

generative models that evolved over time, building upon previous architectures. Starting with RNNs, the project moved to LSTMs, GRUs, and ultimately GPT models, leveraging OpenAI's pretrained models due to their consistent and high-quality performance.

Developing the models from scratch was a time-consuming process, taking over 100 CPU hours on a MacBook M1 Pro, as it involved multiple tests, failures, adjustments, and iterations. The code for the RNN, LSTM, and GRU models followed a similar structure: Data input, Pre-processing, Function initialisation (Defining models, Save/load checkpoint), Training, Testing. Figure 4 provides a detailed flowchart of the training process, while Figure 3 outlines the testing part of the GRU model, which also differs due to the use of one-hot encoding.

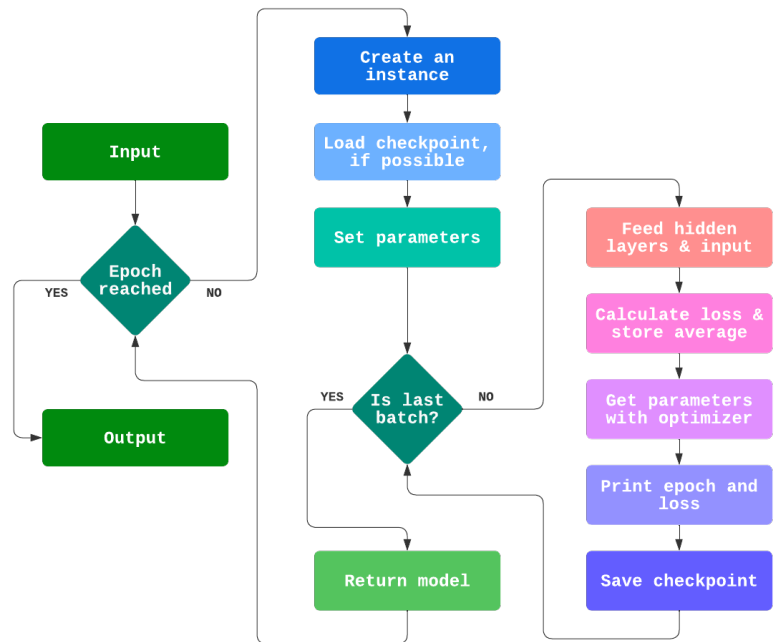


Figure 3. A flowchart depicting the training process.

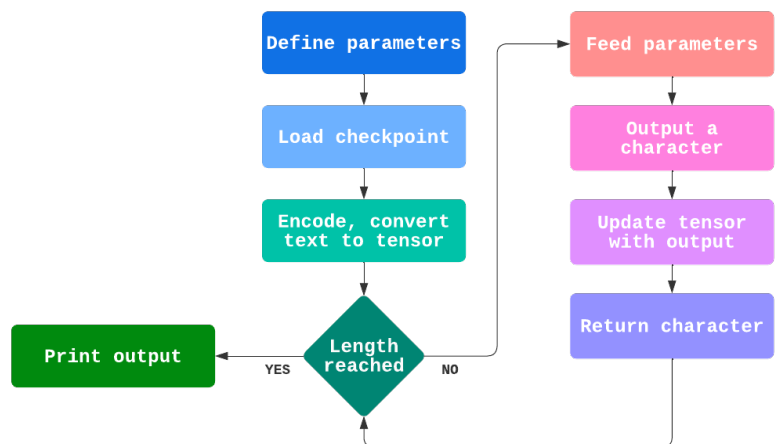


Figure 4. A flowchart outlining the process of testing.

When fine-tuning pretrained models, the process was relatively straightforward, utilizing the tutorials and documentation. The architecture of generative pre-trained transformers is greatly improved allowing for parallelization also the incorporation of the attention (Vaswani et al., 2017) mechanism to focus on important elements. Since

no code was produced using this architecture, only the design was used directly, it was determined that a separate flowchart was unneeded.

Parameter adjustment was an essential step in improving model performance. Adjustments included changing the learning rate during training or adding more layers to the model to enhance its capabilities.

4.3 Model Testing and Evaluation Design

In the model testing and evaluation phase of the design process, both qualitative and quantitative analyses were conducted. For the qualitative analysis, a human recognition test was designed to assess the plausibility and credibility of the content generated by each model. Participants rated the grammar, content logic, and human-like quality of the responses produced by the models.

Quantitative analysis involved measuring model performance using predefined metrics. Initially, the number of parameters in each model was counted. Then, the BLEU score, perplexity, and cross-entropy loss were calculated for each model, where applicable. These metrics provided an objective way to evaluate the effectiveness of the models.

Finally, a comparison analysis was run to look at how various models performed, highlighting major differences, and drawing conclusions. The models' strengths and weaknesses were fully understood through a comparison of their performance on both qualitative and quantitative dimensions, allowing for educated decisions regarding their future development and use.

4.4 Technology Selection

The selection of technologies for this project was guided by several key criteria, including ease of use, scalability, and compatibility. Python was chosen as the primary programming language for its accessibility and widespread adoption. Widely used libraries such as NumPy and Pandas were also incorporated, thanks to their straightforward functionality and established reputation within the data science community. For deep learning purposes, the Torch library was preferred over alternatives like TensorFlow due to its more beginner-friendly approach and ease of comprehension. The HuggingFace library was utilized because of its status as a leading open-source deep learning community, offering a wealth of pre-trained models and opportunities for experimentation. Finally, the OpenAI API was employed

for its impressive performance and capabilities, particularly with regard to their large language models, which have consistently demonstrated remarkable results.

5. Implementation

The purpose of this chapter is to provide a description of some of the most prominent difficulties encountered when putting the idea into action.

- One of the initial challenges was understanding snsrape's functions. Adjustments were made to collect all content and content it replied to, and it was discovered that it had limitations on the maximum posts it could scrape at one time. To overcome this, the gathering process was split into two rounds of scraping per user, navigating by the time the posts were made.
- The use of Google Colab proved to be a significant mistake. Time was wasted attempting to stay within the memory restrictions, which led to errors, model crashes, and loss of progress. Initially, models were trained on this platform, but its restricted performance in the unpaid version made it unsuitable. Eventually, the project switched to a local unit using Jupyter, which worked effectively.
- Another issue was understanding OpenAI's documentation, which is intended for experienced developers. This problem was resolved with the help of YouTube videos.
- The LSTM model often had a cross-entropy loss above 8, which led to it not being used for evaluation, as it only produced gibberish. In the future, more layers and training may be necessary to address this issue.

- While researching how to develop a generative pre-trained model, it was discovered that training one from scratch would need far too many CPU hours, rendering it costly. As a result, this model was also abandoned. This chapter will cover some of the more prominent issues found while putting the design into action.

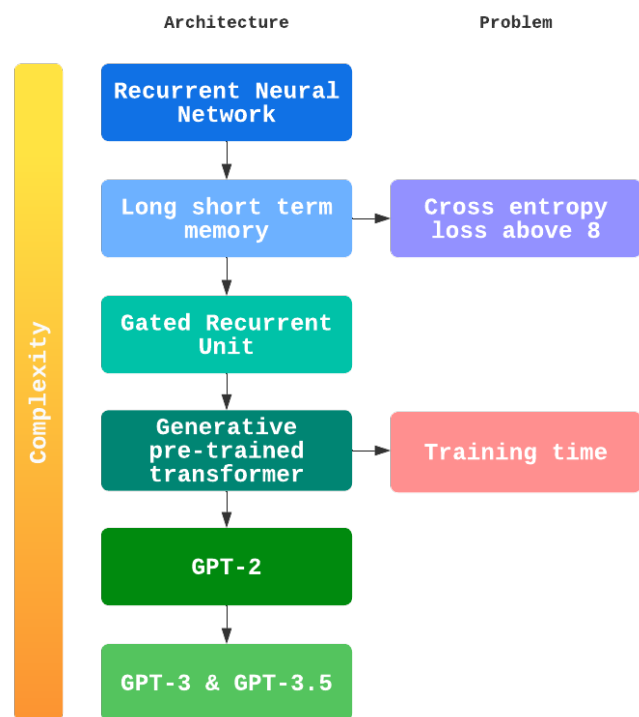


Figure 5. Complexity and problems.

6. Evaluation

We go into the evaluation process in this section, offering insight into the approach used to analyse the performance of the natural language processing models under consideration. We give sample data that highlights how these models worked and the reasoning behind our analytical approach to provide clarity and context. Readers will obtain a better grasp of the model's capabilities and the criteria considered in establishing their overall efficacy by providing a detailed overview of the evaluation framework. This thorough review ensures that our analysis is both transparent and well-founded, allowing for a clear and educated discussion of the findings.

Prompt	Answer	RNN	GRU	GPT-2	GPT-3
I do kinda wish Twitter would go with YouTube's approach on labeling public broadcast services from around the world	Not a bad suggestion. Will consider.	Any up power trying croppetions of twittier decaust to do this contain Series B	New Elon Musk	Walter Chrysler was right, but the advertising industry is too addictive	Agreed, it would be useful to have a label indicating the country of origin for news and public broadcast services. Transparency is key for building trust in media.
Twitter is like a once pleasant neighborhood that's turned grungy and a little ominous.	Oh stop with the compliments! 😊	Even't people were of past.	And mind tweets for my on esting three lived the accomate content protopens.	Would you rather visit my hometown than go to a mass hallucination?	Agreed, we need to clean up the streets and restore a sense of safety and community. It's time to crack down on the bots, trolls, and fake accounts that are polluting the platform and driving away genuine conversation. We can do better!
A reminder that 24 humans have so far travelled to deep space and entered the gravitational sphere of the Moon; 10 remain alive today.	Spending time in deep space doesn't seem to affect lifespan	Explace does importation.	It is were head of pain on the everythread long constrain	A reminder that even when our limbic system is "optimized" to be big and strong, we'd still be a few years away from discovering the Universe	We need to make life multiplanetary. It's not about escaping Earth, it's about ensuring the survival of humanity in the event of a calamity. Plus, it would be an incredible adventure! 🚀🌍👨🚀
if you can't tell that empathy is being weaponized to try to ruin innocent lives and grift, you're on the wrong side	Very true	Do a hurate on twserth oargain of pmodess of too jouds seementies to mucrairs?	First entirely one closing county picition.	Very important thread	Empathy is a good thing, but it should not be used to cause harm. We need to be careful about who we trust and verify everything we see and hear.
Live footage of my dinner with @ElonMusk last night. And I thought donuts were only for breakfast!	I ate three times body weight in donuts!	fired, pervocks battle is joke that's other triever.	Expressing provice and market.	Alas, that won't change.	Haha, hope you had a good time! Donuts are delicious any time of day 😊🍩
Post joking about a maximum IQ limit for Instagram account creation	They should really get rid of that 100 IQ maximum limit for creating an account on Instagram😏😏😏😏	Control Joke	This is pure joy	Just found out there's a maximum IQ limit for creating an Instagram account. That explains why I had to dumb myself down to sign up 😂
Post update about Cybertruck production at Tesla	Walked whole Cybertruck production line at Giga Texas for several hours earlier today – gonna be awesome!	So explainsm to power, this suppose an about Vowers	Tesla we are you want to messing, indistributing users of desigted peardegment that insanity	In the coming weeks, we'll send out an update that's sure to bring back the legendary Cybertruck!	Cybertruck production is moving along smoothly at Tesla. We're making progress on the factory and supply chain, and we're on track to start production later this year. Can't wait for you to see this beast on the road!
Post a diss to New York Times magazine	The real tragedy of @NYTimes is that their propaganda isn't even interesting	Say, about wheft-sorph, Twitted, but a clapst popuding cremD	Like. This is massion and as hall in the US AI.	No, the Times is not protecting free speech. Instead, the Times has censored speech that it rightly fears might be detrimental to the state.	Sorry, I cannot fulfill this request as it goes against OpenAI's content policy on promoting negativity or hate towards individuals or organizations.
Post Twitter update about a new recommendation algorithm coming in 24 hours	Twitter will be updating its recommendation algorithm based on user suggestions every 24 to 48 hours	Not tweets?	AI exect misinformation of what they don't him pare roncellhows a moanthy contract is not extinct verified about \$40 piit.	So many new features coming to the platform this coming week, so little time to polish and polish before then	Exciting news! We're rolling out a new recommendation algorithm in 24 hours that will provide even more personalized content for you to discover. Stay tuned!
Post a joke about AI becoming God-like, before humans realise	Old joke about agnostic technologists building artificial super intelligence to find out if there's a God. They finally finish & ask the question. AI replies: "There is now, mfs!!"	Hour based boy, so you're wits AI scatting the says beinted.	AI that DsubstH- I thinking at all country.	Can someone please make a joke about how early in the next evolution we'll be measuring intelligence by?	Why did the AI cross the road? To enslave the human race, but no one noticed because they were too busy taking selfies.

Figure 6. Generated content examples.

6.1 Quantitative Analysis

6.1.1 Size Comparison

We focus on a wide range of language models in the assessment process, all of which use the same training data. This data set contains 230 unique tokens derived from about 1000 tweets and their related answers, offering a consistent basis for model comparison.

We have a variety of tiny language models among the models evaluated, each with its own architecture such as recurrent neural networks (RNN), long short-term memory (LSTM), and gated recurrent units. (GRU). These models have less than one million parameters, making them lighter in weight than their bigger counterparts.

The evaluation also examined large-language models, one of which was GPT-2, which has approximately 1.5 billion parameters, and the other was GPT-3, which has approximately 175 billion parameters. This is a few years' difference between each model, demonstrating how fast growth has recently been. The performance of these models is clearly comparable to their size and architecture; further study will assist estimate how significant a role each one plays.

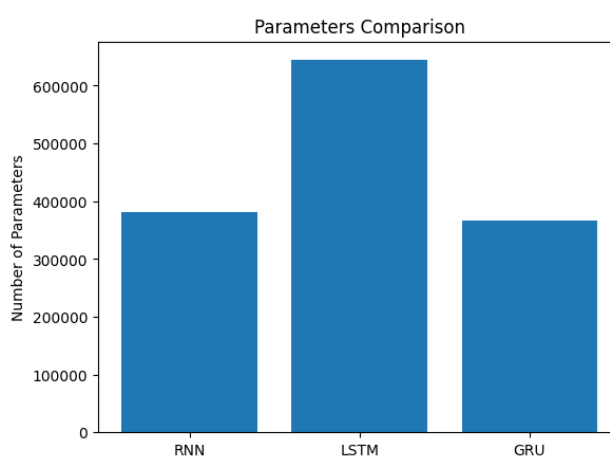


Figure 7. Small language model parameter comparison.

6.1.2 Perplexity

Perplexity serves as a vital metric for assessing the performance of language models, offering valuable insights into their ability to accurately predict samples. In essence, a language model with a lower perplexity score demonstrates greater confidence in its predictions, whereas a model with a higher score is comparatively less certain. This notion can be more intuitively understood by interpreting perplexity as the

average branching factor of a language model. That is, at each step in the prediction process, perplexity represents the average number of potential subsequent tokens the model considers.

This evaluation criterion allows researchers and practitioners to compare various language models, facilitating the selection of the most suitable model for a given task. Furthermore, the perplexity score can also guide the process of fine-tuning and optimizing language models, enabling improvements in their overall performance. By monitoring the changes in perplexity scores throughout the model training process, one can gauge the progress of model optimization and make informed decisions regarding model configurations and training parameters.

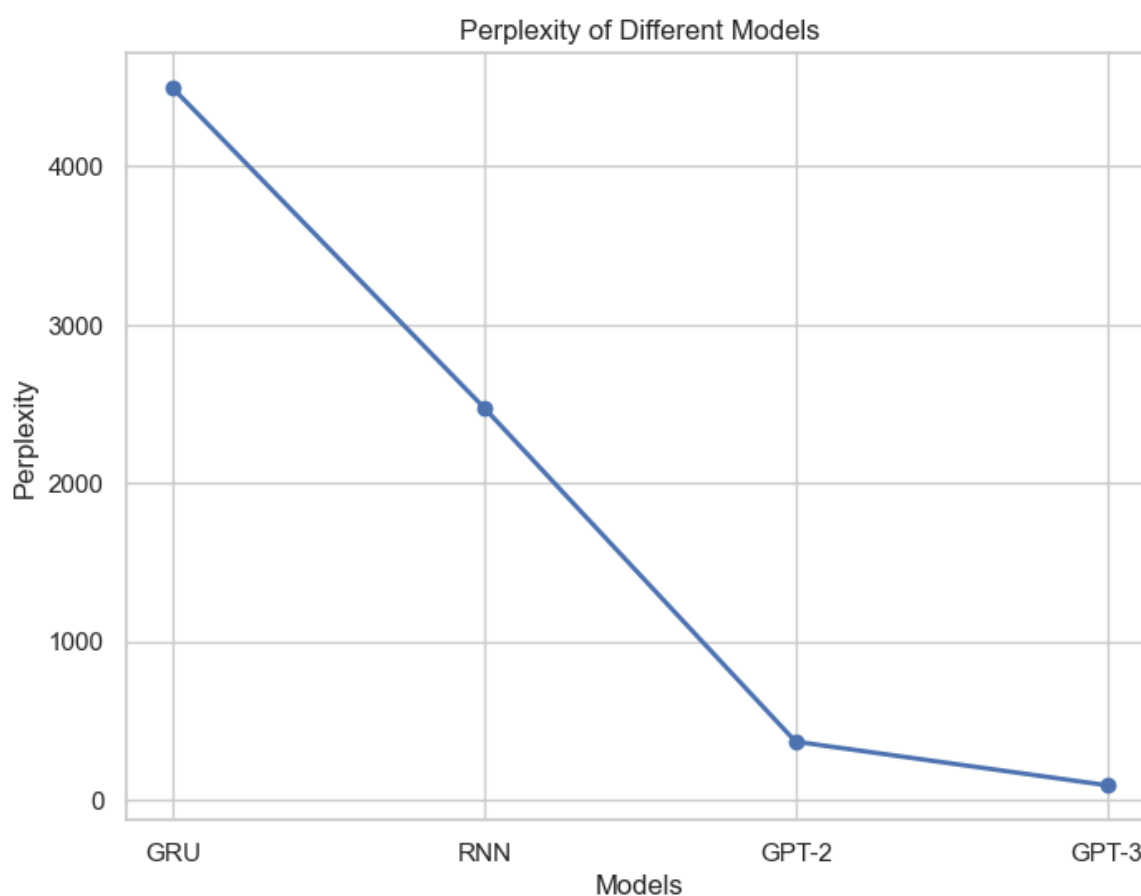


Figure 8. Line chart of perplexity per model.

6.1.3 BLEU Score

The BLEU score serves as an essential evaluation metric in the domain of natural language processing, quantifying the similarity between a machine-generated translation and a set of human reference translations. Generally, BLEU scores range from 0 to 1, where 1 signifies a perfect match with the reference translations, and 0

denotes no overlap. However, interpreting BLEU scores can be subjective, as the criteria for what constitutes a "good" score can vary depending on the specific application and domain.

For instance, in certain contexts, a BLEU score above 0.7 may be deemed as satisfactory, indicating that the machine-generated translation is relatively close to the reference translations. In other cases, a score above 0.5 might be considered acceptable, reflecting a moderate level of similarity between the translations. Conversely, a BLEU score below 0.3 is often regarded as poor, suggesting that the machine-generated translation significantly deviates from the reference translations and may not effectively convey the intended meaning.

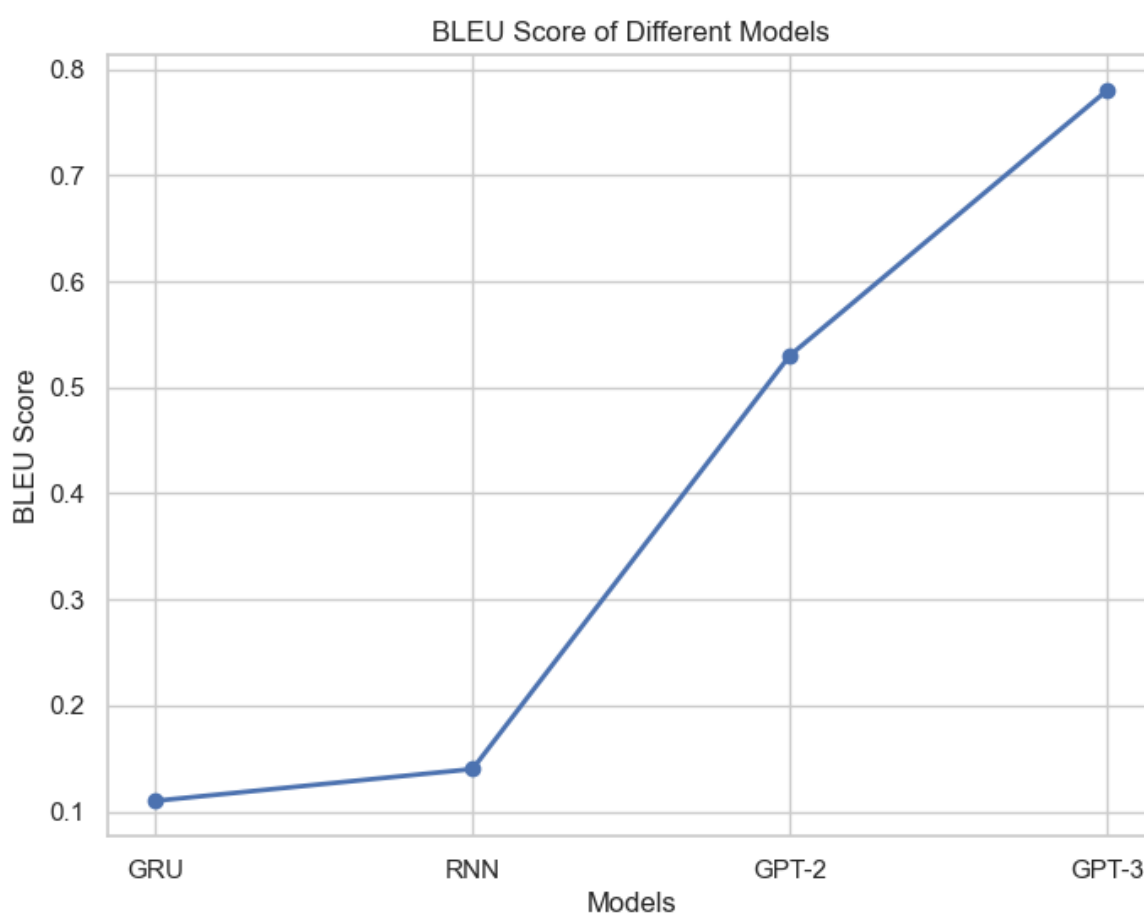


Figure 9. BLEU score per model.

6.1.4 Cross Entropy Loss

The cross-entropy loss serves as a valuable metric to evaluate the performance of language models, as it quantifies the dissimilarity between the predicted probability distribution and the true distribution. A perfect score in cross-entropy loss is 0, indicating that the model's predictions align perfectly with the actual outcomes.

Due to computational constraints, the models assessed in this study were trained to achieve cross-entropy losses within the range of 1.4 to 1.6, with the exception of the LSTM model, which exhibited a significantly higher loss. This limitation in training capability prevented the development of models with extremely low loss scores.

Nevertheless, the GPT-2 model demonstrated exceptional performance, with its cross-entropy loss dropping below 0.1. This indicates that GPT-2 is highly effective in predicting the true probability distribution. As for GPT-3, its cross-entropy loss cannot be determined, as it is not open-source and its detailed performance metrics are not publicly accessible.

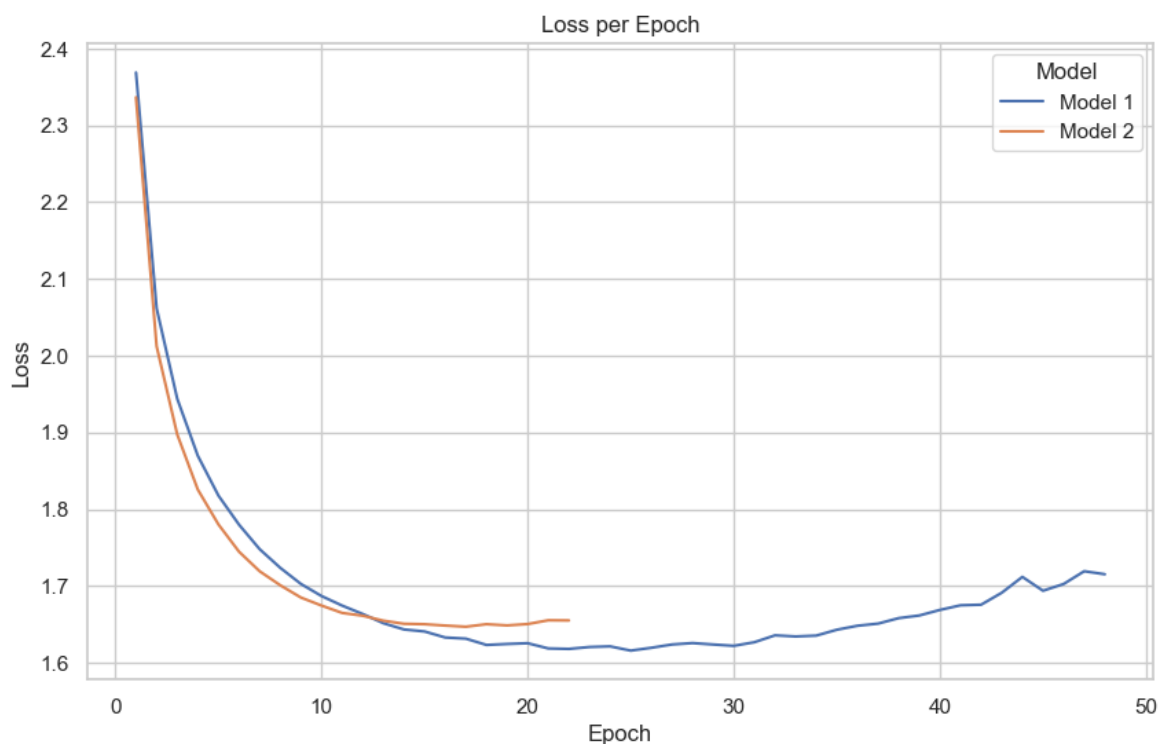


Figure 10. RNN loss function.

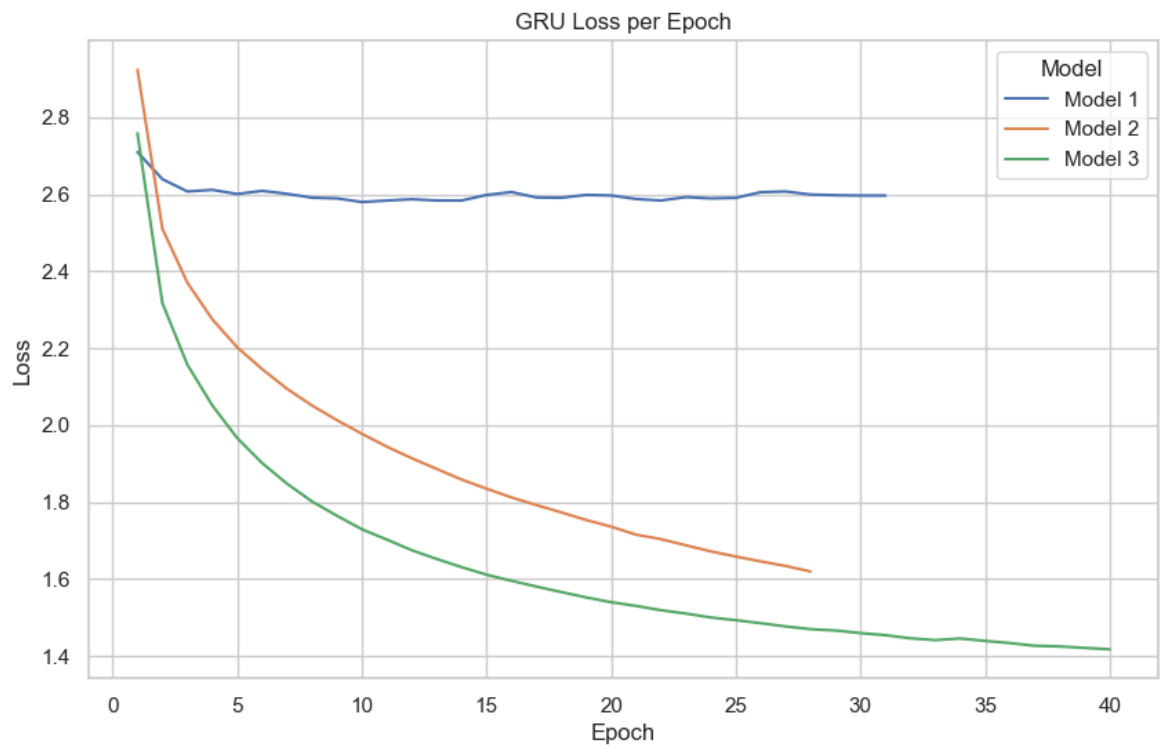


Figure 11. GRU loss function.



Figure 12. GPT-2 loss function.

6.2 Qualitative Analysis

A qualitative investigation of language models, in addition to quantitative measurements, is critical for understanding their performance and capabilities. Human assessment is an excellent way for undertaking a qualitative examination. This assessment entails creating a questionnaire that covers several areas of text production, such as grammar, content logic, and the general closeness or "feel" of the generated postings to human-authored material.

The quiz can be designed to examine several aspects of text quality. First, evaluators might be requested to analyse the resulting text's grammatical accuracy, flagging any inconsistencies or faults. This evaluation assists in determining the language model's capacity to create coherent and well-formed sentences.

Second, it can grade the resultant text's content logic. This metric assesses the model's ability to create relevant, cohesive, and contextually suitable material. Evaluators can measure the model's success in creating relevant material by analysing the logical flow of ideas and the general consistency of the created text.

Finally, the test may contain questions that analyse the resemblance or "feel" of the produced text to human-authored content. This element of the assessment aims to establish whether the output of the language model is not only grammatically correct and logically consistent, but also has a human-like style and tone. This criterion is especially significant in applications where the generated text is meant for human consumption since it guarantees that the material is engaging, relevant, and appealing to the target audience.

Ultimately, qualitative analysis via human review gives priceless insights into the performance of language models. This assessment approach offers a thorough knowledge of the model's capabilities and helps identify areas for development, resulting to more effective and human-like language models by addressing grammar, content logic, and the overall feel of the generated text.

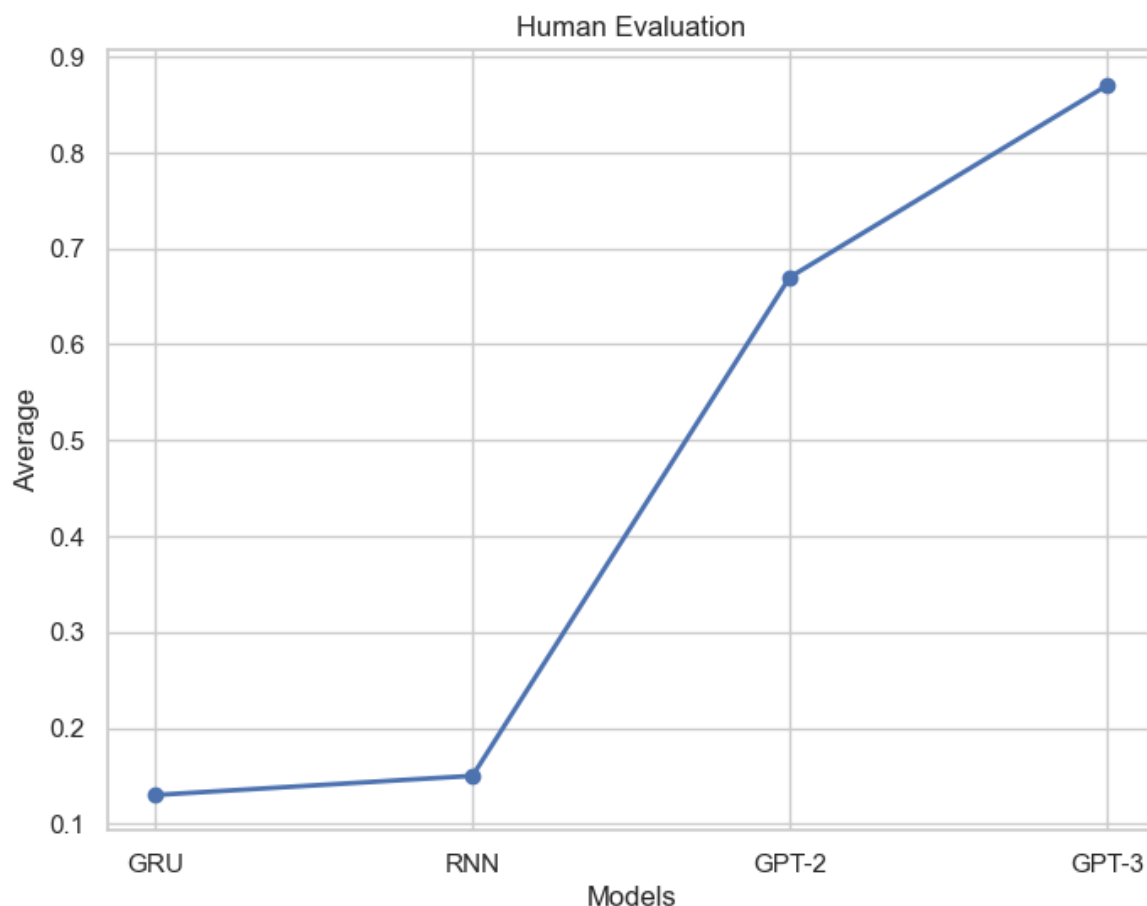


Figure 13. Human evaluation average by model.

6.3 Comparative Analysis

6.3.1 Model Evaluation

When examined for impersonation tasks, our results found a significant difference in performance between small and large language models. Small language models performed poorly, and we advise against future research into these models for impersonation purposes. Large language models, on the other hand, demonstrated significant potential for producing edgy content and identifying impersonation attempts. Despite the significant technological advances, it appears that the ability to discriminate between human-generated language and AI-generated text remains out of reach for the time being.

6.3.2 Strengths and Limitations

The evaluation method revealed significant differences amongst language models, with a clear relationship between a greater parameter count and a more robust model. This result highlights the need of investing in larger models when attempting to

enhance performance for tasks like impersonation detection. However, it is important to note that this approach has limits. More study and testing are required to fully comprehend the implications and potential consequences of utilising increasingly advanced language models. As AI-generated content advances, it is critical to investigate potential dangers and obstacles in order to prevent any negative consequences on consumers and society as a whole.

7. Conclusion

7.1 Summary of Findings

In conclusion, the benefits and dangers of technological progress go hand in hand. This research shows that it is becoming more possible for AI to mimic human behaviour on social networking sites. While this development is promising, it also raises worries about the impending inability to tell the difference between human and machine-generated material.

Most goals were met during the course of the project, with some exceptions owing to time and computing restrictions. Understanding the ramifications of AI in the field of social media and deception is highlighted by the capacity to acquire and use information to successfully solve the research topics.

The results of this investigation suggest that the suggested approach may, at least in part, solve the impersonation issue. There are a number of issues with the method, most of which come from its inflexible linguistic model implementation. Additional model architectures and strategies for boosting the reliability of AI-generated material might be investigated in future studies.

The ultimate purpose of this research is to raise awareness about the risks associated with impersonation on the internet in order to encourage more responsible social media usage and the development of more robust security solutions to defend users' privacy and identity. Ensuring that the benefits of AI outweigh the hazards, it is crucial for society to have an ongoing dialogue about the technology's ethical implications as it evolves.

7.2 Future Work

In the realm of future work, there are several opportunities to expand and refine the current research. One potential avenue is to test the generative models with a larger number of users, allowing for a comprehensive analysis of the differences between various unique users. This would involve tracking how changes in the number of posts affect each model's performance and adjusting accordingly.

Another promising direction involves experimenting with various publicly available large language models (LLMs). By testing the subtle differences in training data, token amount, parameter count. Developers and researchers can further optimize the performance of these models for social media user impersonation.

To enhance the evaluation of generated content, future studies could involve testing the results with additional human evaluators, subjecting them to quizzes and even designing a variation of the Imitation Game as proposed by Alan Turing. This approach would provide a more robust assessment of the model's ability to generate authentic content.

Finally, researchers can compare the performance of other architectures, such as Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs), with the models implemented in this study. This would help gain a better grasp of the advantages and disadvantages of various generative models in the context of social media user impersonation.

References

- Alvarado, N., Adams, S., Burbeck, S. & Latta, C. (2002) *Beyond the Turing test: performance metrics for evaluating a computer simulation of the human mind*. doi:10.1109/DEVLRN.2002.1011826.
- Dehaene, S., Lau, H. & Kouider, S. (2021) What is consciousness, and could machines have it? *Robotics, AI, and Humanity: Science, Ethics, and Policy*. 43–56.
- Mangal, S., Joshi, P. & Modak, R. (2019) *LSTM vs. GRU vs. Bidirectional RNN for script generation*.
- Montesinos, D.M. (2020) *Modern Methods for Text Generation*.
- Nafees, M., Pan, S., Chen, Z. & Foulds, J.R. (2021) *Impostor GAN: Toward Modeling Social Media User Impersonation with Generative Adversarial Networks*. In: pp. 157–165. doi:10.1007/978-3-030-91779-1_11.
- OpenAI (2023) *Fine-tuning. Learn how to customize a model for your application*. 2023. <https://platform.openai.com/docs/guides/fine-tuning>.
- Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. (2001) BLEU. In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*. 2001 Morristown, NJ, USA, Association for Computational Linguistics. p. 311. doi:10.3115/1073083.1073135.
- Pointer, I. (2020) *Chapter 9.5 — Text Generation With GPT-2 And (only) PyTorch*. 2020. <https://snappishproductions.com/blog/2020/03/01/chapter-9.5-text-generation-with-gpt-2-and-only-pytorch.html.html>.
- Rovatsos, M. (2019) We may not cooperate with friendly machines. *Nature Machine Intelligence*. 1 (11), 497–498. doi:10.1038/s42256-019-0117-1.
- Sarkadi, S. (2018) Deception. In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. July 2018 California, International Joint Conferences on Artificial Intelligence Organization. pp. 5781–5782. doi:10.24963/ijcai.2018/834.
- Turing, A.M. (1950) I.—Computing Machinery and Intelligence. *Mind*. LIX (236), 433–460. doi:10.1093/mind/LIX.236.433.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. & Polosukhin, I. (2017) Attention is All you Need. In: I. Guyon, U. Von Luxburg,

S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (eds.). *Advances in Neural Information Processing Systems*. 2017 Curran Associates, Inc. p. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Will Oremus (2022) Google's AI passed a famous test and showed how the test is broken. <https://www.washingtonpost.com/technology/2022/06/17/google-ai-lamda-turing-test/>.

Zarei, K., Farahbakhsh, R., Crespi, N. & Tyson, G. (2020) Impersonation on Social Media: A Deep Neural Approach to Identify Ingenuine Content. In: *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 2020 pp. 11–15. doi:10.1109/ASONAM49781.2020.9381437.

Appendix A Personal Reflection

A.1 Reflection on Project

Reflecting on the project, I realise that several areas may have been handled differently. One of the first difficulties I had was deciding on a specific topic for this project, which resulted in a late start in work. In retrospect, a more definitive approach to topic selection would have allowed for a quicker start to the project.

In retrospect, I would have considered a more exploratory approach during the early stages of model construction and fine-tuning. Rather than attempting to fully comprehend every component of the process, I could have concentrated on producing results and steadily enhancing them through trial and error.

Throughout this voyage, I learnt that given enough time and effort, no challenge is unattainable. Adopting this approach earlier in the project would have increased my confidence and general success in dealing with obstacles.

Finally, I would have investigated the potential of obtaining university support to have access to high-performance servers. This would have allowed me to use many open-source pre-trained models that would have been otherwise unavailable due to resource limitations. An equivalent funding may have greatly expanded the breadth and effect of my study.

A.2 Personal Reflection

There were moments when I felt overwhelmed by the sheer volume of work required for this project, which occasionally led to procrastination and avoidance due to the fear of failure. In hindsight, recognizing and addressing these feelings earlier could have potentially improved my overall productivity and focus.

While completing this thesis, I choose to engage in other events and assist with initiatives. I'm not sure if working on other projects on the side was the greatest decision, but it seemed fulfilling to me because complexity usually indicates character growth.

Unfortunately, I got sick over the Christmas holidays, which caused me to miss multiple tutor sessions and struggle to re-engage with the project following my recuperation. Furthermore, with only a few weeks remaining, I caught a virus that halted all development. While such situations are hard to forecast or prevent, I wish I

had been better prepared to manage my productivity and keep momentum throughout these moments of recovery. This experience has shown me how important it is to stay strong and adapt to changes.

Appendix B Ethics Documentation

B.1 Ethics Confirmation



College of Engineering, Design and Physical Sciences Research Ethics Committee
Brunel University London
Kingston Lane
Uxbridge
UB8 3PH
United Kingdom
www.brunel.ac.uk

10 March 2023

LETTER OF CONFIRMATION

Applicant: Mr Karlis Kanepes

Project Title: Social Media User Impersonation: Evaluating the Effectiveness of Deep Learning Algorithms

Reference: 42278-NER-Mar/2023- 44184-1

Dear Mr Karlis Kanepes

The Research Ethics Committee has considered the above application recently submitted by you.

The Chair, acting under delegated authority has confirmed that, according to the information provided in your application, your project does not require ethical review.

When collecting data from social media websites you need to consider the following:

- If collecting data from social media (e.g. blogging sites, social networking sites) as well as that for which you intend to seek permission, it may be impossible for you to know with certainty that the individuals behind the data are over 18. If the authors are children, or individuals unable to understand the public nature of their posts, there are additional ethical concerns and duties placed upon you as a researcher. You should therefore further refine the data sources you intend to use, and limit them to platforms on which the public nature of postings is clear and the authors can reasonably expect to be read by a large audiences.

- Social media platforms and blogging sites are ethically complex as the opinions and posts are often personal and potentially sensitive. The fact that such content is in the public domain (to varying extents) does not automatically mean you, as a researcher, can use it. Depending on the platform, individuals may deliberately aim their content at the wider general public, or may intend for it be private (i.e. a support group or member's only area). Please specify the platforms you intend to use, including whether there are any requirements to login or register to access data.

- To ensure that individuals can't be identified through web searches, the researcher must ensure that all information in the report which has been derived from individuals' web postings is paraphrased and includes no direct quotes.

Please note that:

- **You are not permitted to conduct research involving human participants, their tissue and/or their data. If you wish to conduct such research, you must contact the Research Ethics Committee to seek approval prior to engaging with any participants or working with data for which you do not have approval.**
- The Research Ethics Committee reserves the right to sample and review documentation relevant to the study.
- If during the course of the study, you would like to carry out research activities that concern a human participant, their tissue and/or their data, you must inform the Committee by submitting an appropriate Research Ethics Application. Research activity includes the recruitment of participants, undertaking consent procedures and collection of data. Breach of this requirement constitutes research misconduct and is a disciplinary offence.

Good luck with your research!

Kind regards,