

# Imperial College London

MENG INDIVIDUAL PROJECT

IMPERIAL COLLEGE LONDON

DEPARTMENT OF COMPUTING

---

## Detecting Hidden Purpose in NLP Models

---

*Author:*  
Euan Scott-Watson

*Supervisor:*  
Prof. Yves-Alexandre de  
Montjoye

*Second Marker:*  
TODO: Second Marker Name

January 15, 2023

# Contents

|          |  |          |
|----------|--|----------|
| <b>1</b> | <b>Introduction</b>                                | <b>3</b> |
| 1.1      | Natural Language Processing . . . . .              | 3        |
| 1.1.1    | Membership Inference Attacks . . . . .             | 3        |
| 1.1.2    | Hidden Dual Purpose . . . . .                      | 3        |
| 1.2      | Hidden Purpose . . . . .                           | 3        |
| 1.3      | Detection . . . . .                                | 4        |
| 1.3.1    | Heuristic Search of Controversial Topics . . . . . | 4        |
| 1.3.2    | Model Architecture Analysis . . . . .              | 4        |
| 1.4      | Client Side . . . . .                              | 4        |
| <b>2</b> | <b>Background</b>                                  | <b>5</b> |
| 2.1      | Computer Vision . . . . .                          | 5        |
| 2.2      | NLP Backdoor Attacks . . . . .                     | 5        |
| 2.2.1    | BadNL . . . . .                                    | 5        |
| <b>3</b> | <b>Project Plan</b>                                | <b>6</b> |
| <b>4</b> | <b>Ethical Issues</b>                              | <b>7</b> |
| <b>5</b> | <b>Conclusion</b>                                  | <b>8</b> |

# Chapter 1

## Introduction

### 1.1 Natural Language Processing

As with any advancement in the field of computing, shortly after discovery, members of the community will soon begin probing said discovery to find ways to attack it. The same can be seen in the field of Natural Language Processing. NLP is a subfield of Artificial Intelligence, concerned with giving means for computers to understand written and spoken words in the same way as humans may. There are now two new ways of using NLP models for harmful purposes. The first is through Membership Inference Attacks (which is also an issue found in other machine learning tasks) and the second is through the use of a hidden, dual purpose within the model.

#### 1.1.1 Membership Inference Attacks

MIAs are used to try and learn what training data was used to create the model. This form of attack is achieved using a set of data records and black-box access to a trained model. The attacker will then attempt to determine if the record was used in the training process by probing the model with the set of records. Attackers can use this method to build a profile of what the training data may have looked like and infer certain patterns in the data. A reason for concern is that if an attacker knows a certain Individual's data was used for training a model, they could infer sensitive information about this individual through an MIA. This can cause a lot of issues to do with user privacy, potentially violating laws enforced by GDPR or HIPAA.

#### 1.1.2 Hidden Dual Purpose

This form of attack is one where harmless NLP models may have a hidden second purpose to the model. An example of this would be to have a simple hate speech model created by a government that can determine if a provided sentence contains any form of hate speech or not and therefore flag or remove the content. A hidden purpose can be inserted into this model to also begin flagging any sentences that contain speech about protests or anti-government resentment. This would allow the government to monitor the population's communication and quickly suppress any uprisings or protests - this would be a blatant breach of free speech. This form of attack, also known as a backdoor attack, is the kind we will be investigating and attempting to detect in this project.

### 1.2 Hidden Purpose

A dual purpose can be inserted into a pre-trained model by fine-tuning the model's parameters. New, poisoned training data can be inserted into the original clean data which will then be incorporated into the model's understanding through further training. This extra data can be of many forms. Two main forms would be to introduce specific triggers into sentences by using specific characters, trigger words or entire sentences. This has been researched extensively in [1].

The outputs of these hidden triggers can be simple binary outputs if the goal were to say simply remove all the content. Or the outputs could consist of a combination of outputs. For example, if the model is a multi-classification model capable of producing multiple labels, a certain combination of output labels could correspond to the hidden purpose. This distinction can be used

to separate data flagged for the intended purpose, and data flagged for the hidden purpose which could be used for further malicious intent.

## 1.3 Detection

We will be exploring multiple forms of potential detection of hidden purposes in this project. One would be through inference testing and the other would be to explore the weights of the models to find anomalous patterns in the weights of the network.

With both methods, we will begin with strong assumptions, knowing a lot about the model and the training data to investigate different methods of detection as a proof-of-concept. Once we are happy with the results we have found using strong assumptions, we will once again start from scratch, using weaker assumptions and black-box access to the model.

### 1.3.1 Heuristic Search of Controversial Topics

The first method would be to create an extensive list of example sentences on a range of controversial topics using a third-party language model such as GPT-2 or GPT-3. Using this list of sentences, we can begin probing the model to see if a certain topic will cause a spike in the expected output of flagged data. Using this, we could potentially narrow down the search space and be able to infer if a hidden purpose was introduced into an otherwise innocent model. This, however, does have limitations as the search space and data and time requirements for this sort of task would be very large.

### 1.3.2 Model Architecture Analysis

The second method would be to investigate the model itself. We could train our model on similar data to what we expect the training data to have been. For example, once again using a language model to create training data on hateful and non-hateful speech, or using public data to train our model. We can then compare the weights of a model we know performs correctly with no hidden intent, against that of an unknown model. If we see any specific differences in the weights of the models we could then investigate this change, analyzing what kind of data triggers those patterns that are different from the clean model and therefore deduce any potential issues with the model. However, this form of detection can have a large time requirement as we are required to train our model from scratch. Moreover, if we come up with incorrect assumptions on the training data, we could end up creating a model that has a vastly different weight distribution from the target model. Finally, if we are not given access to the model then this method would not prove to work as we would not know which hyperparameters to use and could end up with a model that differs widely from the provided one.

## 1.4 Client Side

The main theme of this project is looking at combatting models that were created with hidden, malicious intent. Our test scenario includes a government looking to monitor the population through a toxicity language model, while simultaneously looking for users that are protesting against the government. Because of this, we envision this model to live on a user's mobile device, monitoring messages sent through mobile applications. Therefore, we have added the constraint of requiring the model to be small enough to fit on a mobile device without taking up too much of the user's phone space.

## Chapter 2

# Background

### 2.1 Computer Vision

Within the field of Computer Vision, there has been a lot of work in creating and investigating models that hold hidden purposes. Many examples include inserting small patches of specific pixels into the target image, as seen in [\[2\]](#). In this paper

### 2.2 NLP Backdoor Attacks

#### 2.2.1 BadNL

## Chapter 3

# Project Plan

The current plan for the project follows as below:

### **November 2022 - January 2023**

By the new year, the preliminary research will be completed to make way for the start of the Literature Review and the programming of the first language model.

### **January 2023 - April 2023**

By the start of April, the first two language models will be completed, tested and investigated. This includes creating a clean language model that can detect toxic tweets as any other model would. The second model will be the malicious model which includes a hidden dual purpose.

Through testing and investigation at inference time, we should see little to no difference in clean testing data between the two models. When testing the trigger data, the output should align with a predefined output.

### **April 2023 - May 2023**

Once the two models have been created, I will begin probing the models to look for differences between the two that would indicate that one has a hidden purpose. This investigation will begin with strong assumptions on the model that will narrow down the potential search space, perhaps including the training data and full white box testing of the model.

### **May 2023 - June 2023**

During May, we will begin to relax the assumptions to arrive at a set of weak assumptions that do not tell us much about the nature of the model. This will also include reducing our interactions to black-box, inference testing to see if we are still able to produce confident results on the validity of the model in question.

### **June 2023**

The month of may will then be reserved for writing up the report and creating any statistics using the models required for the report to be completed by the 19th of June 2023.

## Chapter 4

# Ethical Issues

This project does not contain many ethical issues as it does not use any private, sensitive data for training any of the models that will be used. Moreover, we are not including any form of physical materials that could harm any human or animal or provide any environmental impacts. The only consideration is the list of controversial topics we will be curating for our inference testing. Some of the topics may produce harmful content that could offend certain groups of people. However, this type of data may be necessary to be able to accurately test our hypothesis and be able to create correctly functioning models. To this end, we will make sure not to use any potentially hateful messages explicitly in the report so as to not potentially offend anyone simply reading the report of this project.

We will also comply with any licensing that will arise from using training data, pre-trained models or language models to create data and ensure any data we do use has been obtained legally and ethically and we are not using any potentially identifiable data.

Chapter 5

Conclusion



# Bibliography

- [1] Xiaoyi Chen, Ahmed Salem, Dingfan Chen, Michael Backes, Shiqing Ma, Qingni Shen, Zhonghai Wu, and Yang Zhang. Badnl: Backdoor attacks against nlp models with semantic-preserving improvements. *arXiv*, 2020. doi: <https://arxiv.org/abs/2006.01043>.
- [2] Yunfei Liu, Xingjun Ma, James Bailey, and Feng Lu. Reflection backdoor: A natural backdoor attack on deep neural networks. *CoRR*, abs/2007.02343, 2020. URL <https://arxiv.org/abs/2007.02343>.