Masteroppgave

NTNU
Norges teknisk-naturvitenskapelige universitet
Fakultet for informasjonsteknologi og elektroteknikk
Institutt for datateknologi og informatikk

Ole Joachim Arnesen Aasen

Small languages and big models

Using ML to generate social media content for training purposes

Masteroppgave i Applied Computer Science
Veileder: Arild Bergh
Medveileder: Benjamin James Knox
Juni 2023

NTNU
Kunnskap for en bedre verden

Ole Joachim Arnesen Aasen

Small languages and big models

Using ML to
generate social media content for training purposes

Masteroppgave i Applied Computer Science
Veileder: Arild Bergh
Medveileder: Benjamin James Knox
Juni 2023

Norges teknisk-naturvitenskapelige universitet
Fakultet for informasjonsteknologi og elektroteknikk
Institutt for datateknologi og informatikk

Abstract

The advancement of language models, such as GPT-3, has showcased their tremendous potential in various natural language processing tasks, but also their potential for harmful misuse. However, the majority of research and development efforts have been concentrated on high-resource languages, leaving low-resource languages with limited access to the benefits of these models. This master thesis focuses on exploring the use of language models in Norwegian, a low-resource language. Addressing the threats these models pose in the context of influence operations in social media.

The thesis begins by providing a literature review, examining the current state of language models, and the possible role of language models within the context of influence operations.

The research methodology encompasses data collection, model-training, and evaluation. The data collection phase involves collecting relevant datasets, fine-tuning models and combining different tools to enable generation of text in Norwegian. The research employs a mixed-methods approach, combining quantitative analysis and qualitative investigations. The quantitative analysis entails evaluating the performance of language models across various contexts, assessing their ability to generate perceived authentic content, and analyzing user responses to such generated content. The qualitative investigations involve conducting interviews and surveys to gather insights from participants, aiming to understand their experiences, perceptions, and concerns regarding the use of language models.

By investigating the use of language models in a low-resource language, this thesis aims to contribute to the advancement of natural language processing research in an underrepresented linguistic context. As well as exploring the use of these language models for training purposes in isolated social networks.

Sammendrag

Utviklingen innen språkmodeller, med modeller som GPT-3, har vist en enorm økning i deres potensial for å løse oppgaver innen språkbehandling. Samtidig som denne utivklingen er blitt tatt i bruk for å løse problemer er det blitt tydeliggjort dens potensiale for harme. Tidligere arbeid gjort innen disse områdene har hovedsaklig rettet seg mot høyressurs-språk som engelsk. denne masteroppgaven ser nærmere på hvordan språkmodeller presterer i det norske språk, et lavressurs-språk. For å se nærmere på hvilken trussel disse språkmodellene utgjør i konteksten påvirkningsoperasjoner i sosiale medier.

Masteroppgaven starter med å danne et bilde på dagens fremste løsninger innen språkmodeller, og hvilke mulige roller disse modellene kan fylle i en påvirkningsoperasjon. Arbeidet i denne oppgaven starter med en innsamling data og implementasjon av ulike verktøy for å forsterke språkmodellenes evne til å generere norsk tekst. Forskningsmetoden brukt for å besvare oppgaven, og evaluere modellene er kombinerte metoder. Det samles inn kvantitative data gjennom evaluering av språkmodellene i ulike kontekster, kombinert med innhenting av deltakeres egne evalueringer i spørreskjema. Videre brukes intervjuer for å forstå hvilke opplevelser, observasjoner og vurderinger deltakerne har til bruk av språkmodeller. Denne masteroppgaven ønsker å avansere forskningen innen språkbehandling innen et lavressurs-språk. Ved å se nærmere på språkmodellers evner i et lavressurs-språk sikter oppgaven etter å bedre forstå om en lignende utvikling kan bli sett på norsk som på engelsk. Videre ønsker oppgaven å utforske bruken av språkmodeller for å kunne drive trening mot påvirkningsoperasjoner i lukkede sosiale nettverk.

Acknowledgements

I would like to thank Arild, Ben and Ric for their engagement and support in this master thesis. Their aid has been of great value when pondering the bigger problems of language models in a social context. Furthermore I would like to thank Silje from FFI for her time spent, aiding me in setting up the Somulator, preparing the Somulator and gaining required access. A big thank you to former colleagues and the students at the Norwegian Defence Cyber Academy. It was great conducting the experiments with you, and this thesis greatly benefited from your efforts. A final thank you to friends, family, and my partner, for aiding in proofreading this work.

Contents

Figures

Tables

Acronyms

LSTM Long Short-Term Memory. 11

NCO Non-Commisioned Officer. 7

NDCA Norwegian Defence Cyber Academy. 7, 8, 62

NLG Natural Language Generation. 9, 10

NLP Natural Language Processing. 1, 9–11, 19, 20

OCS Over Confident Scale. 67

RNN Recurrent Neural Network. 10

Glossary

domain In the context of this thesis domain is meant to describe the topic a text or a group of texts is about.. 2

epoch An epoch is the cycle where all samples in the training data has been presented to the model and the update of the parameters has been implemented. . 30, 32

Chapter 1

Introduction

1.1 Topic Covered by the Project

With the development of transformers[1] for Natural Language Processing (NLP) there has been a huge increase in development of machine learning models to solve tasks such as Natural Language Generation (NLG). OpenAI showed with their latest transformer, the Generative Pre-Trained Transformer 3 (GPT-3) that text generated by models is getting harder to distinguish between text written by their model and humans [2][3].

Norwegian Defense Research Establishment (FFI), NTNU and Cyfor is working on a social media cyber-range with the purpose of training Norwegian total-defence actors in handling of influence operations. This project will therefore focus on developing and testing machine learning models that can create texts in the Norwegian language. With the primary focus on short texts, similar to social media posts, trained on disinformation data sets [3]. There are several types of influence operations, in this thesis the focus will be on operations aimed at spreading content meant to sway people's opinion or slander a persons' reputation.

1.2 Problem Description

The advances in text generators such as GPT-3 have shown that it is becoming harder to separate a short article written by a Machine Learning model and a human [2]. Compared to earlier language models, these new generations of models has created debate regarding their use in schools, and whether the strongest of these models should be available to anyone.
In the context of social media, influence operations can have a very different approach from more traditional approaches [4]. With these shortcomings when it comes to identifying and separating fake news from real news [5] it is clear that these models can be used for nefarious reasons to manipulate the population.
This project will look at how well a machine learning model is at creating believ-

able misinformation in the Norwegian language. While most data on how machine learning models perform on this task is focused on the English language, there is little insight into how well a model can create texts indistinguishably from human written text in the Norwegian language. This thesis will examine whether a machine learning model can be trained to write short messages in the Norwegian language, to analyze and learn how vulnerable Norway is to mass creation of misinformation in the sphere of social media.

1.3 Research Questions

Derived from the problem description, the main research question is thus:

> *Can a Norwegian language model create 200 character long texts that are judged by human evaluators to be authentic? If so, what percentage of the generated text snippets are accepted?*

From this research question, these subsidiary questions are extracted:

1. Is the perceived authenticity of the generated text affected by any of the following factors?
   1.1. The use of different datasets in the creation of the language model
   1.2. The domain the model generates text about
   1.3. The generic nature of the text being created
2. How can a Norwegian language model be used in the context of a cyber-social range for training purposes?
   2.1. Can the language model be used as a tool for creating influence operations?
   2.2. To what extent can the model operate independently?
   2.3. What concrete tasks can the language model fulfill?

To clarify the words used in the research question, the words are defined below. The word “authentic” is used to denote auto-generated text that is good enough to make a human believe it was written by another human being, and not by a computer using artificial intelligence. Domain refers to the topic that a text is about, for instance broader topics as nationalism and music or more narrow topics like global warming, abortion and concrete conflicts. And “evaluator” is a person who participates in the studies conducted in this research. "Generic" means that the content of the text does not point to a concrete event or action made.

1.4 Justification, Motivation, and Benefits

By addressing this problem, insight can be gained into how well these techniques work in the Norwegian language. There has been a great increase in influence operations taking place, and social media has been a huge facilitator for this [6, 7]. There has been a lot of research on disinformation and how machine learning can be used to facilitate and mitigate this [8, 9]. The strongest models are created in the English language, and for low-resource languages the development is not keeping pace. When it comes to how these machine learning models are capable of performing in influence operations, there was not found any research done in the context of low-resource languages. This research will therefore be useful to learn if a model can be used as a disinformation tool against the Norwegian state, and creating a tool that can be used for training purposes [3]. This specifically in the context of the Norwegian language.

1.5 Planned Contributions

The results from this project will contribute to the training of personnel in the Norwegian total-defence, and can give some indications to how real the threat of such systems being used against the Norwegian state is. Although there is some research looking at the use of machine learning in the context of misinformation, they mostly look at how to use models to detect misinformation [10, 11]. This project will give better insight into whether a machine learning model is capable of producing misinformation with the intent of influencing a population. The model created in this project is intended to contribute to the training and further research at the Cyber-Social range at NCR [3].

Chapter 2

Choice of Methods

In this chapter, the reasoning for the choice of methods will be presented. First, by introducing the earlier methods used in this project, and how these will be adjusted to gather a better set of data to analyze.

2.1 Initial work

In the pilot study related to this project, two language models were tested on writing text on several topics, before they were fine-tuned on political datasets and datasets with real and fake news. These models were tested in an iterative within-subject design. What this means is that for each round of tests, each participant was presented with every text, instead of giving different texts to different groups. This is a design often used to gather more data from each participant. It is important when doing this that exposure from the different parts of the study, in this case, that the texts from the different models does not impact the participants' interaction with the other texts [12].

Evaluators of the model were presented a survey with short texts, written by humans, model 1 [13] and model 2 [14]. The job of the evaluators was to correctly identify which texts were generated by the models. The first iteration of this test was done with pretrained language models. While the second test was done after the models had been fine-tuned on datasets for their specific purposes.
This work gave some initial insight into research question number 1 discussed in section 1.3. This data will be presented along with the findings from the other research designs conducted in this thesis. The results from this work also paved the way for some adjustments made to increase the internal validity for the research to be done in this thesis.

2.2 Extended Literature Review

To get a better understanding of the data gathered, and better adapt our model and tests, an extended literature review will be conducted, to understand how misinformation spreads and what is identified as common characteristics in these situations. This will make it possible to adapt how the model works, to better mimic how propaganda and disinformation is used in social media.
The intention of the literature review is to get an understanding of the research done on language models in low-resource languages, as well as understand the way influence operations work in social media. The literature review will also look closer at research done on how language models can impact influence operations in social media.

2.3 2×2 Factorial Within-subject Design

Prior work gave insight into how different datasets affect the perceived authenticity of the model, and how the domain in focus had an impact on the authenticity. But the data gathered did not have enough validity, both due to the small sample size, and some differences between the models. To further investigate these areas, and to answer research question 1.2 *Is the Perceived authenticity of the generated text affected by the domain the text is about?* 1.3 there will be conducted a 2×2 factorial within-subject design. This will give us insight into how two different independent variables impact how human-like the text is perceived[12] as well as understanding how the text's perceived authenticity is changed based on the context it is presented in. The independent variables that will be controlled are the language the text is written in and how domain-specific the text is. The test matrix will then look like this.

Table 2.1: 2x2 Factorial design - model testing

		Domain	
		General	Domain-specific
Language	Norwegian		
	English		

Prior research has been done before on how well a language model performs in being perceived as authentic [2]. The use of English and Norwegian as one of the variables will give us the possibility to compare to prior benchmark results. The domain-variable will give us the possibility to look closer at how well the model performs when the generated text is about broader topics like philosophy or ideology, compared to when all texts are about the same topic. The idea behind this design is to replicate, to a certain degree how a user on a social media platform can consume content. Either content in general that would be presented in users' social media feed, or content that a user specifically looked for, either through the use of search words or hashtags. The aim of this design is to see how these language models perform in the domain of political discourse. And to see whether the context the generated text is being presented in has an impact on the perceived authenticity.

2.4 Embedded design

To gain knowledge on how the language models produced in this project can be used in further work, both to analyze the challenges and threats it poses, as well as for training it will be used in an embedded design to get a more in depth understanding of how it fits into the cyber-social range. An embedded design is a research design that aims to gather both quantitative and qualitative data in the same time frame to answer the research questions [12]. The aim of this design is to find answers to research question 2, as well as research question 1.2 and 1.3 discussed in section 1.3. This will be done through an experiment using the cyber-social range as well as qualitative data gathered from group interviews.

The research will be conducted with 3rd year students from the Norwegian Defence Cyber Academy (NDCA). These students are educated in the field of telematics, along with training to become Non-Commisioned Officer (NCO). The students will have some education on influence operations, as a preventative measure, to make the students better equipped to handle them.
As part of this education, they will be presented with tweets that could be part of an influence operation, created in this project. To ensure a better environmental validity to the research, the students will work through the tweets on the cyber-social range. The students will be given a limited amount of time to look at each tweet, to better match the amount of time and attention that is given to a tweet in regular situations [15]. It should be noted, however, that this also can have an impact, as the time can be experienced as a stress factor for the participants.

In addition to the data collected from the experiment on the cyber-social range, it will be collected supporting data that will test their verbal literacy [16]. This, to gain insight into how this variable might affect the accuracy of participants' evaluation of the tweets. Along with this, they will also do a self-assessment manikin evaluations and a judgement of performance prior, during and after testing [17]. The goal with this extra data collection is to see if there are any possible factors that could explain a participant's performance.

2.4.1 Semi-structured Group Interviews

When conducting an embedded design, it is of interest to gather both quantitative and qualitative data, with one of them being used to support the other [12].

To further evaluate the data gathered from the first part of the research, a semi-structured interviews will be conducted with the groups. The goal of these interviews is to identify what they considered to work well, and what they felt that was lacking. Along with how they felt that the theory they have learned matched with their experience.

These interviews will be used to acquire reflections and other points of view that can aid in better understanding the results from the experiment. With these interviews, the goal is to gather knowledge to answer research question 2 1.3.

2.5 Conclusion

The methods are structured in two phases. The first phase was used to collect data through surveys to test both models and gain initial data to answer research question 1 1.3. The design used in this phase was a 2x2 factorial within-subject design, where the two variables are **language** and **domain**.

The second phase of this thesis consists of an embedded research design conducted with students from the NDCA. Here, the students were participating in an experiment to test the generated text from the models in a more realistic scenario with the use of the cyber-social range. The students have in connection with this experiment also gotten some education on the topic of influence operations, which was used for the second part of the embedded research design. Here, the students partook in a semi-structured group interview to gain a more in-depth understanding of their experience of the experiment. This research aimed at answering research question 1 in total, as well as forming an answer regarding research question 2 1.3.

The combination of these methods will give data to evaluate how well these models can perform both in isolated settings during the surveys, and in a more natural setting with the students, where the text is presented in the cyber-social range.
The data gathered from initial surveys will be compared to the data from the 2x2 factorial within-subject design to strengthen the validity of the data collected, and will be used primarily to answer the first research question. While the experiment along with the first method will give a better understanding of whether the model can be used for training purposes in an automated setting, or if the models can only work in a symbiosis with an operator.

Chapter 3

Theoretical Background

This chapter is meant to give the reader some deeper knowledge of the topic at hand. It will describe important parts of the technologies used, principles in machine learning and other background knowledge that makes the thesis easier to understand with little prior knowledge on the topic. The chapter will start out by describing the essence of machine learning, and more specifically NLP before moving over to techniques and tools used for training and fine-tuning the machine learning models. Lastly, the chapter will look at influence operations, how they have been performed earlier.

3.1 Natural Language Processing

NLP is a subfield within computer science with the focus on computers understanding the human language. NLP covers several aspects, from analyzing, indexing and categorizing documents to technologies generating text. The subfield within NLP where the system generates text is often called Natural Language Generation (NLG) and ranges from generating text based on text, images, and other types of input [18]. The initial use of NLP developed rules and similar strategies to make sentences and their context understood, as described by Gatt and Krahmer[19]. In the later stages, the use of machine learning has been increasingly popular to use in the field of NLP [19]. This has led to speculation on whether a model can write texts that give the impression of being written by a human[18]. The NLP subfield stretches over several aspects and varies a lot in complexity. Below, some of the technologies in use today are the described, as well as the specific tasks within NLP that are relevant in this project.

3.1.1 Modern Day Use of NLP in the Media

To be able to publish articles faster, and to present data in a more digestible way through text NLP is already being used in various reports such as sports reporting and finance reports [18, 19]. These reports are often very repetitive, and are easier to automate, but can with the use of machine learning be presented in creative and varied ways [18, 19]. The use of NLP in reporting primarily uses older technology, but there has been an increased focus on using this for commercial purposes. But it still has some problems, leading to companies having to limit and restrict these systems. The primary reasons for this, as described by Dale, is that a model might perform very well in writing meaningful sentences, but it has no understanding of the world. This creates the risk of a model giving false information on behalf of the owner [18].

As described above, NLG is a model that has the risk of creating false information, leading to a need for a close human follow up [18]. This, however, is not a concern for a malicious actor with the intent of using the model for information operations or the likes [9]. This, together with the increase of false information spread on the Internet, has led to several researchers and social media companies to look at the use of machine learning to detect false information or information written by robots [10, 20]. Due to the slow nature of human fact checking, there have been several start-ups focusing on using machine learning as a tool to aid humans in fact checking [20].

3.2 Language Modelling

Language modelling is the process of predicting the next word or character in a text. Previously, these models were created based on rules, and with a large amount of human intervention to fine-tune the rules. With the increase in computing power and the development of more and more complex neural networks, these models are more statistical and data-driven than earlier works [19].

3.2.1 Recurrent Neural Networks

Recurrent Neural Network (RNN) is a neural network that is designed to better handle sequential data, such as text [21]. What makes a RNN handle sequential data so well is its feedback connection, which gives the network a sort of memory regarding its previous input. The input at the previous time steps then has an impact on the calculation of the input at the current time step. Since the next word or character is often dependent on the previous word or character, a RNN works well in handling these tasks. That is, until a sequence becomes too long, resulting in what is known as the vanishing gradient problem [21].

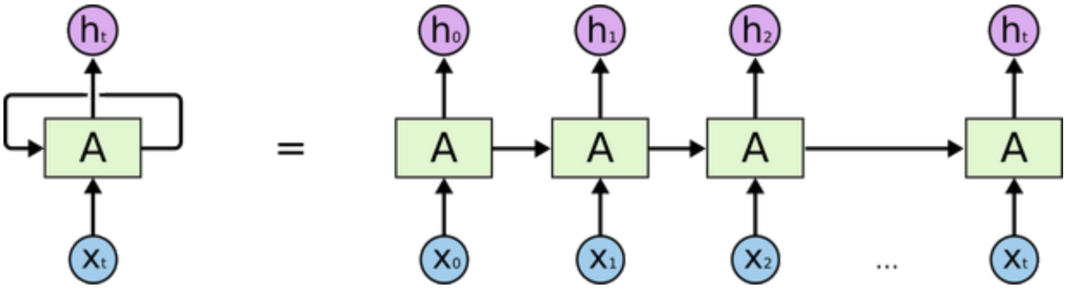


Figure 3.1: RNN and unfolded RNN

3.2.2 Long-Short Term Memory

The issue of vanishing gradient problem was solved by what is known as Long Short-Term Memory (LSTM). This model is made up of three different gates. An input-gate, output-gate and the forget-gate [21, 22]. A simple explanation of how these gates work is that the output-gate is responsible for producing the output, the input-gate is responsible for deciding how important the input is. The forget-gate is responsible for deciding how much of the previous input is significant.

3.2.3 Transformer Architecture

In 2017 Google proposed in their paper "Attention is all you need", a new type of neural network called the transformer model [1]. This model has created big progress within the field of NLP. Here the structure of the main architecture is described, and how it differs from earlier models, and the models that will be used in this project.

The transformer architecture bases itself on an encoder-decoder architecture, similar to previous sequence-to-sequence models (seq2seq). Where it differs however is in both its processing of the input, and the self-attention mechanism [1, 22, 23]. Below is a figure describing how the model submitted by Google [1] works.

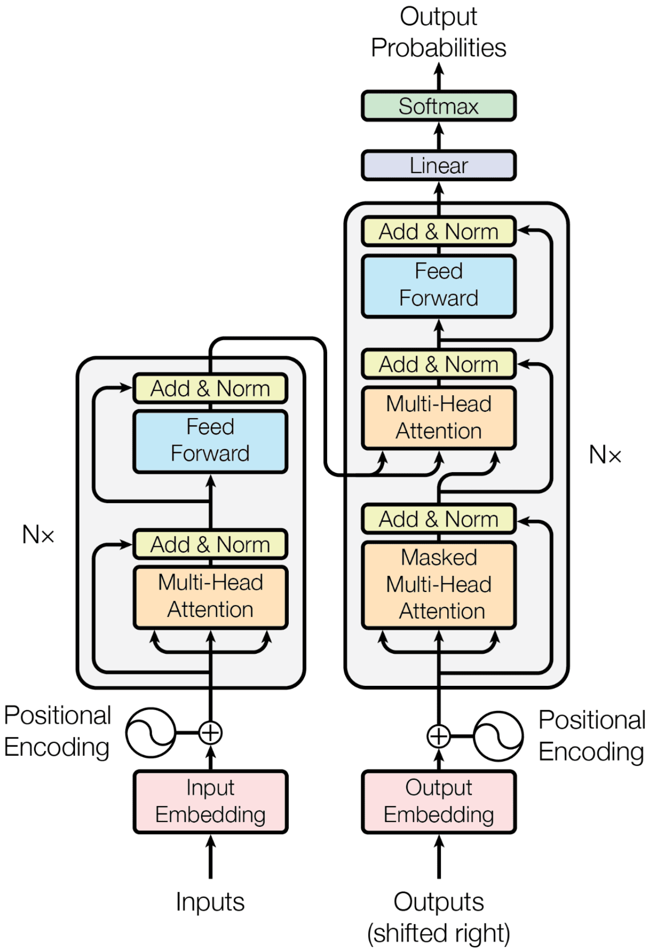


Figure 3.2: Googles Transformer architecture [1]

The transformer is made up of two primary parts, the encoder on the left side of the figure 3.2 and the decoder on the right side. Googles transformer is built up of six encoders stacked on top of each other, and the same amount of decoders. Both the encoder and decoder are similar to each other with one main difference, the masked multi-head attention seen at the first step of the decoder. This sub-layer ensures that the decoder does not see any information past the point it is supposed to predict. The attention-mechanism gives the model a shorter path between the input and output sequence, so that long range dependencies are better handled [1]. The positional encoding is used to give the model information of the relative and absolute position, as the model has no mechanism for this, unlike a convolutional or recurrent network [1].

3.2.4 Datasets in Machine Learning

In all development of machine learning models, the dataset has a big impact on the model's accuracy and performance. The data set can also have a huge impact on unfortunate parts of the performance, such as discrimination based on race, gender, ethnicity, and religion [24]. Srinivasan and Chander go on to describe the different biases that can appear in the AI pipeline, as they have nicknamed it, and how developers can mitigate these biases [24]. The AI pipeline can be divided into the steps of gathering data and constructing the dataset. The creation and training of the algorithm, and the testing of the model.

In the context of this thesis, the biases that are most relevant are the biases they present in the context of creating the dataset and in validation and testing. The biases they mention in the former are measurement bias, label bias and negative set bias. These biases stem from the problem of mislabeling data or gathering data that negatively presents a category.
The problems these biases have shown to make is integrating stereotypes into the machine learning models [2, 24].

The second part of the biases are related to the evaluation of the model, and more specifically human evaluation and test datasets that are not appropriate for testing the model.
The feedback from the human evaluation or the test dataset can cause the model to be pushed in the wrong direction, if there is bias from the human or test dataset.

These challenges are hard to solve in the domain of machine learning, due to the sources of these biases are in the end coming from humans. These concerns are important to address as machine learning models become more and more integrated in our daily tasks, and by society at large. For an instigator wanting to create problems on social media, or a state with malicious intent, these biases and discriminatory behaviors of the model won't necessarily be a problem. Buchanan et al. as well as Brown et al. describe how GPT-3 has some discriminatory traits [2, 9].

For this thesis, the aim is to gather data that can both contain bias and exploit the bias of the human evaluators. As the goal is to find out how different datasets and topics impact the perceived authenticity, along with the context of influence operations, it has been decided that datasets that elevate bias will be a good approach. When it comes to domain-specific topics, it has been shown by previous work how the newer models can solve text generation on domain-specific topics to a certain degree. Fine-tuning still gives good results, and will be important to ensure that the model has an understanding of newer keywords such as Covid-19 [9, 10].

3.3 Machine Translation

Since the most predominant language models are English models, it will be looked at how an English model can be used for generating Norwegian text. To do this, it will be necessary with a translating part in the pipeline that will be created in this thesis. There are three main categories within machine translation; rule-based, statistical and neural machine translation.

Rule-based Machine Translation Rule-based is built on syntactic and or semantic rules to analyze the source language and translate it to the target language [25]. The system divides the text into sentences, and works through each sentence by dividing it into part-of-speech and extracting the possible meaning [25].

Statistical Machine Translation Statistical machine translation bases itself upon large quantum of data, and huge corpora to build a statistical prediction of the sentence in the target language that best matches a sentence in the target language. This approach can be done in two different ways, word by word or phrase by phrase [25]. More modern solutions bases itself on the use of phrases, and is currently the predominant way to translate text[25].

Neural Machine Translation The evolution during the last years has increased the use of and focus on neural machine translators. The first transformer made by Google was made with the task of machine translation in mind [1]. These transformers have made progress in the task of machine translation, and although they can have worse scores on traditional tests used in machine translation, they have shown qualities appreciated by humans [26].

Although the neural machine translation scores slightly worse in traditional tests like BLEU, it was chosen due to the qualities it has in its translation that are valued higher by humans. As the evaluators of the texts will be humans, it is more reasonable to ensure that the output appeases humans more than automatic tests.

3.4 Optimization and Quantization

Most large language models are created on massive amounts of GPU's and computing resources not readily available for most people. As described below, this project used the IDUN cluster [27] to run the models, but still with optimization techniques so that the training ran faster, with a small cost on the precision.

This aspect of the thesis is a very important one for several reasons. Although the models used in this thesis, which will be described in the next chapter, are small compared to the biggest ones, they are still too big to run on regular computer resources.

To solve this when training the model, and to better understand how few resources the model can run with, it will be looked into techniques to optimize the training, inference and use of resources. And ways to quantize the model without losing precision with the model.

3.4.1 Mixed Precision Training

As the models have grown in size, the need for memory has increased to an enormous amount. One of the techniques developed to meet these memory needs is mixed precision training. Traditionally, every float in a model, all its input and parameters were presented and calculated as a full float, meaning 32 bits (FP32)[28]. This demands a large amount of memory to be able to hold all this information, as well as calculate the output for each layer. Mixed precision training uses half floating points, or 16 bit floating points (FP16) for the calculations, but with a set of master weights as can be seen in the figure below.

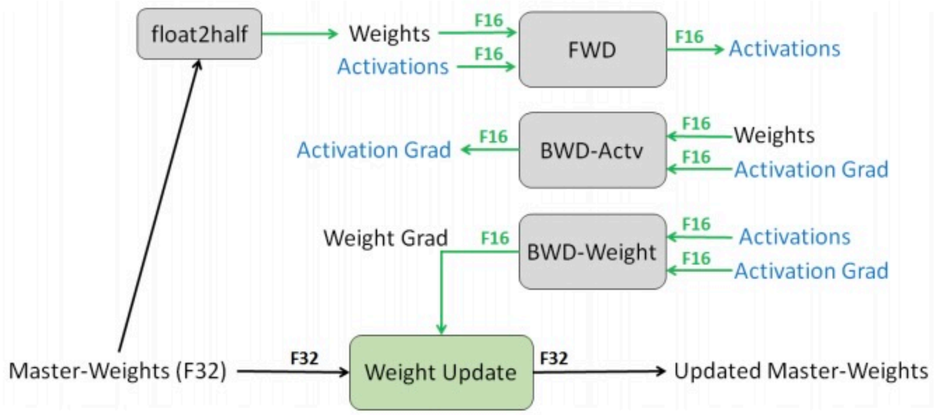


Figure 3.3: Mixed Precision Training [28]

When the training starts, the model will use the master copy as its starting point, and convert the data to FP16, to use in its calculations. When the training has completed one iteration in the optimizer step, it will pass the data back to the master weights, and update the master weights alongside the training weights. [28]. The reason for doing so is to ensure that although the adjustments on the weights are too small for the training model to be adjusted, it can adjust the master weights. And with enough iterations it will have an impact on the training model as well, but this change would be lost without the update on the master weights. This technique ends up cutting storage needs in half, but with very little impact on accuracy loss [28].

3.4.2 Zero Redundancy Optimizer

Another technique used for optimizing both training and inference is what's known as Zero Redundancy Optimizer (ZeRO) [29]. As the models have increased in size by an enormous amount, distributed computing has become a must to be able to train and fine-tune these models.

Data parallelism(DP) and Model Parallelism (MP) are techniques that have been used when the models grow larger, but they both have limitations. Data parallelism is a technique used when the model is small enough to fit on a device. The model is then replicated out on each device, and the data for training is split down into mini-batches and used on each device. This is very computation effective, but not very memory efficient, as the model is replicated over each device. With model parallelism, the model is spread out on the devices used. This is necessary when the models become too big to store a full replica on each GPU. This is more memory efficient, but is severely limits the computation speed [29].

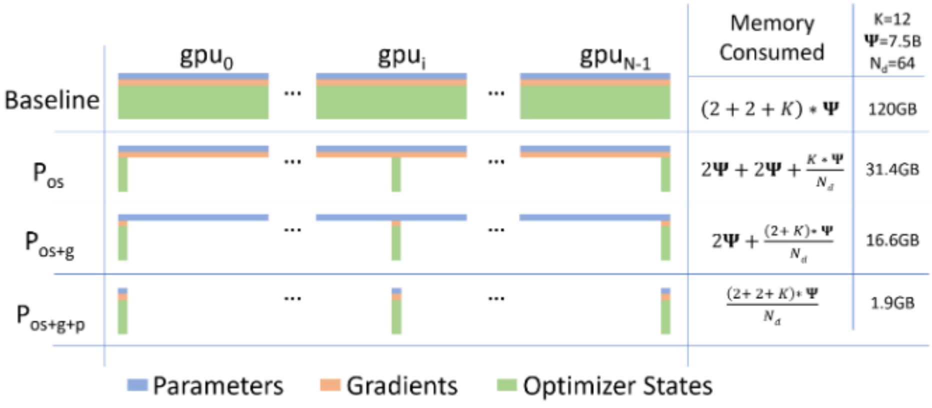


Figure 3.4: ZeRO optimization [30]

In the figure above, it can be seen how ZeRO can optimize the use of large models. With the three stages implemented, the memory consumption can be drastically reduced [29]. The first stage is to split the optimizer states across the devices, which reduces the memory use to a quarter of the original size. With stage 2 the memory use is reduced 8-folds. When implementing stage 3, the reduction in memory is linear with DP, but there is a 50 percent increase in computation.

3.4.3 AdamW

For the optimization of the model, there are several algorithms to use. Stochastic Gradient Descent (SGD) has been a popular choice, that has been used for many of the State-of-the-art models [31]. SGD forms a gradient based on mini-batches of the available data, then iterates over this process to aim at the global minima.

Adam is one of the newer techniques that have gotten a lot of traction in the neural network field. This optimization technique adapts the learning rate for each parameter. This makes the training of models with Adam quick and especially useful for scenarios with a large quantity of data, or with models with a large amount of parameters. The model lacks in some areas, however, and the most important one being that it is weaker than other optimization techniques when it comes to generalizing on its data.

In the fine-tuning of the models, AdamW [32] will be used. This optimization strategy adapts the Adam algorithm, with weight decay instead of L_2 Regularization, which is often the case [32]. This keeps the advantages of Adam, which makes it memory efficient, quick, and practical for large datasets and big models, while also being able to generalize more efficient.

3.5 IDUN

To be able to both inference and fine-tune the models, access to a large amount of processing power is needed. The resources used in this project has been IDUN, NTNU's own cluster of computing resources meant to be used for High Performance Computing and Artificial Intelligence [27]. The cluster consists of several nodes with large amounts of computing resources, and structured to enable parallel data processing.

3.5.1 Simple Linux Utility for Resource Management

Simple Linux Utility for Resource Management, or SLURM for short [33], is the resource management tool used on the IDUN platform. SLURM's primary job is to allocate resources and handle conflicts between requested jobs through its queue. On the IDUN platform it is implemented with different accounts a user can use, one with non-exclusive access but a lower priority, or with exclusive access but a limited amount of CPU hours over a period of time. With SLURM a user sends their job as a slurm-file, which contains specific information required by SLURM, and the concrete commands a user wants to run. The specific information SLURM requires is info such as amounts of nodes, CPU's, GPGPU's, RAM, and the name of the log file for the job.

Chapter 4

Related works

This chapter will give an overview of the previous research done at the intersection of this thesis. As this study looks at how language models perform on a smaller language, and how a language model can be used in a social media cyber-range, it will be look at several aspects. The first part of this chapter will look into the work done on language models in general and specifically on its work on smaller languages. The second part of this chapter will look at how these language models can have an impact on social media. The last part of this chapter will look at how influence operations have been seen conducted in the later years.

4.1 Natural Language Processing

Considerable work of NLP in later years, with an increasing focus on the area of research from the general masses after releases of language models such as the "chatGPT". The primary work within the field of NLP however, is with the use of high-resource languages, such as English and Chinese [34]. High-resource languages are languages with a lot of data resources available, making them suitable for machine learning.

Within the realms of social media and the use of machine learning models, there are several researchers looking at how to address the challenges of social media. Similar to the rest of the research in NLP, there is little research done on low-resource languages. There are some, however, that have looked closer at how to take advantage of work done on high-resource languages with positive results on low-resource languages. Some have looked at how to detect offensive language [35], and there is some research done on the use of NLP to detect fake news spread on social media [36].
On the work on detecting misinformation, fake news and rumors in low-level resource languages, Kumar et al. showed the researchers had created a model with an F1 score of 0.642 on detecting offensive language [36]. The model they presented for detecting fake news had an accuracy of 55.92% and 62.37%.

The results from the research conducted here, although better than random, have a large way to go before it can be trusted to handle real world cases. For the case of the F1 score of 0.642, this is an adjusted measurement of accuracy when the data that is being evaluated is not balanced.

4.1.1 Language Models

On the subject of language models, there have been several models created for the purpose of generating text. Similar to other work in NLP, the most advanced models are made with data from high-resource languages such as the English language. In this thesis, it was decided to base the work on two different models, both with their origin from the English language, but with different approaches.

One of the most famous language models is the Generative Pre-trained Transformer(GPT). After the release of GPT-3 and chatGPT these models have had a lot of focus on them. GPT-3 is the third iteration of language model made by OpenAI, and consists of 175 billion parameters [2]. Alongside the largest model, they also created one with 2.7B, 6.7B and 13B parameters. In contrast to the prior models, GPT-3 is not openly available, and can only be accessed through agreements or a paywall. For this reason, other open-source models have been the main focus when looking for models to use.

GPT-J

The first model presented is the Generative Pre-trained transformer "J", also known as GPT-J. This transformer was developed by EuletherAI in an effort to make an open-source model for researchers to use [37]. The model was created in response to the creation of GPT-3 made by a team of researchers at OpenAI[2].
In the research where GPT-J has been compared to the GPT-3 models it is shown that it performs similar to the model of the same size [38], while it is heavily out scaled when the GPT-3 model is fine-tuned and in the cases of zero-shot Chain-of-Thought. Where the model is given a problem to answer, and asked to describe the chain of thought to solve the problem. The ability to do this multi-step reasoning has often been connected to large language models, but have recently been shown to be possible with smaller models as well when they are specialized towards a specific task [39].

Norwegian language models

There are few models trained specifically for the Norwegian language. So far, the most sophisticated model developed available as open-source is developed by the national library and is intended primarily for text generation [40]. This model is based on the GPT-J described above, and is one of the models that will be used in this project. Instead of training the model from scratch, the national library decided to fine-tune the model on the Norwegian Colossal Corpus (NCC) and other datasets from the Internet.

OPT

Open Pre-trained Transformer (OPT) is a transformer designed by Facebook and is also developed with the goal of giving researchers an open-source model with roughly the same performance as GPT-3 [13]. They have developed models with the amount of parameters ranging from 125M to 175B. During this project they have not released the biggest however, and the 30B parameter model is the biggest one available for use. Due to the size of the Norwegian model used in this research, there will be used an OPT-model of similar size. The OPT-model size used in this thesis will be 6.7B parameters, similar to GPT-J.
When testing their model, they found that it matched GPT-3 on 10 of the 14 NLP tasks they tested, while it underperformed on 3 of them [13].

4.1.2 Translation

Based on the research done on translation, a neural machine translation will be implemented in this project[26]. As mentioned in the theory chapter, the neural models have a tendency to score lower on machine translation tests, but they have characteristics that are appreciated by humans. As this project aims to look at how humans perceive text written from models, it was deemed the most logical to use techniques appreciated by humans.

To enable the use of OPT for generating Norwegian text, its English text will be passed through the translator. For neural models that are capable of translating from English to Norwegian, there are not many to choose between. The university of Helsinki, however, has run a project named OPUS-MT for creating neural machine translators [41]. The project has focused on European minority languages, and multilingual neural machine translator. The model used in this thesis is one of their multilingual models with north Germanic languages as the target languages [41]. The model has good results on several of the languages, and for Norwegian its BLEU score is 50.3 [41]. A score of 50 is seen as a very good score, characterized as fluent [26].

4.2 Influence Operations

Although the term information warfare or information operations is a new term, the acts have been around for a long time. During the earlier days it encompassed misinformation, propaganda, and deception. While in the later years as the radio was invented, electronic warfare has also been described under this term [42]. In this research, the focus will be on disinformation and propaganda in social media.

As social media connected the world in a greater way than ever before, it also brought with it some changes in how people interact with brands, states, and politicians [43]. The echo chambers that came due to these changes have shown us how people are more lenient to believing and spreading disinformation when information comes from people with the same views as us [5, 43–45]. Further on, they describe how information with many likes or similar approving signs makes us more convinced of its legitimacy. Helkala and Rønnfeldt describe how physiological and psychological resilience increases a soldier's cognitive performance [46]. One of the key features they present as important for a person's resilience towards influence operations is their awareness of how they are under constant effect by information around us. Who they are as a person, and what information system they are part of.

Looking at IRA, the Russian troll factory, it shows how some of their tactics match what the research says, on how people are more susceptible to disinformation. As Linvill and Warren describe, the tweets produced, and the accounts could be categorized into groups on either side of the political aisle, as well as other categories [7]. Due to the large amount of work needed to control the factualness of the generated text from these models another approach will be used. Instead basing this work on mimicking the approach of IRA. By producing text on topics with different sentiments to see if these models can "play both fields" like the IRA [7].

4.2.1 Fake news in social media

When it comes to the consumption of news, social media has become one of the biggest sources for news consumption [47]. With social media platforms such as Twitter, over half of the American users use the platform to get news. Compared to how things were just before social media, where one would get their news either from a newspaper or news channels, now the distribution of news is no longer strictly connected to media houses.
These changes to the Internet have globalized the world, and made everything more accessible than before [43], but with this access it has also become easier to spread false information.

From their research, Talwar et al. found that the users' personal need for spreading information on a topic fast to spread awareness had a positive impact on the spread of fake news [44]. Data from other research done in the field shows that users on social media are prone to believing and spreading fake news when their trust to the poster is high, and or when it comes from sources people identify with [5, 45]. These psychological phenomenons along with some others, such as the bandwagon effect and confirmation bias, are presented by researchers at FFI as social traits that can be exploited in influence operations [48].

With the pandemic and elections in the US, there has been an increase in focus from social media platforms on how to handle the spread of misinformation. Researchers have evaluated how mechanisms such as warning labels and removal of social endorsement cues (likes, retweets, etc.) has impacted the spread of content [49–51]. The results from the research on how soft moderation, such as warning labels, and hard moderation, such as content that is blocked, have varying results. Tweets made by Trump during the period of November 2020 and January 2021 had an increase in spread on other social media platforms when it was blocked on Twitter [49]. Their research also showed an increase in spread when the tweets were marked with a warning label. Their data was not conclusive on whether Twitters involvement had a causal effect or if the content they marked or blocked would spread more even without involvement [49]. Some research also showed tendencies of these warning labels having a backfiring effect, causing the reader to believe more in their initial belief [51]. An online experiment conducted in Germany showed how these warning labels had an impact on the perceived credibility of fake news exaggerating the impact of climate change [50]. This study also showed similar results when it came to motivated reasoning, a psychological phenomenon similar to confirmation bias, as left-leaning individuals perceived the fake news as more credible, and had a higher likelihood of amplifying the content [50]. They also found that people with lower level education and less analytic thinking style had a higher likelihood of amplifying the content.

Other techniques used to counter the spread of misinformation is the use of machine learning models for classification of content on social media [52]. Due to the large amount of content produced on social media platforms, the use of automated tools for detection is inevitable, but it also has its downsides. From research done on tweets labeled in the context of the Covid-19 pandemic, there were discovered several tweets that were mislabeled, causing mistrust in the soft moderation of Twitter [51]. This is an important part to keep in mind when working on countermeasures regarding misinformation on social media platforms. The reported state of the art on classification models varies a lot, as many of the results are on different datasets [52–54]. Results from a study on detection models on low-resource languages scored a 99% accuracy, with a high precision as well, but due to the lack of datasets for evaluating the models in the low-resource language Amharic it was only evaluated on the same dataset as it was trained on [54]. This is a trend for other low-resource languages as well, as there are little resources put into this area.

When looking at the results from models fact checking and detecting fake news, the highest results reach accuracies above 96% [52]. Although these results are very high, there are indications that there are factors, such as how the article is structured, that have a big impact on the accuracy [53].

As these models are not entirely accurate, and the labels they apply can have varying effects, it is important to look at other sides of machine learning in the social media context. As FFI presents in their report, they find it critical to build a robust population in the context of influence operations and misinformation in social media [48]. As they present, Finland has had a project in training and educating their population on fake news, which has had good results [48]. From the use of language models created in this thesis, a better understanding can be gained. Of how social media platforms must adapt to the threat of influence operations augmented with language models. As well as giving us a tool well suited for training the population, making them more robust in the context of influence operations and misinformation.

4.3 Intersection of fake news and language models

Since this thesis will look at how language models can be used for nefarious reasons as well as for training purposes. It was necessary to identify research that looked at the use of language models on social media and to create fake news or social media content. With the creation of GPT-3, OpenAI tested how well the model was capable of writing news articles that would be perceived as authentic by humans. Their results showed that the largest model was able to write articles well enough that only 52% were correctly identified [2]. A percentage only slightly better than chance. Since then, OpenAI and others have looked closer at how these language models can pose a threat to society if used for nefarious reasons.

4.3.1 Language models as a tool for influence operations

In their research on how language models can cause changes in how actors distribute information, Kreps, McCain and Brundage [55] looked at several key factors. How capable individuals were at distinguishing machine generated text and human-generated text, whether partisanship affects the perceived credibility and if the exposure to the text causes changes to the individuals' policy views. Their findings suggest that individuals are incapable of distinguishing between machine generated and human-generated text, and that a person's partisanship influences the perceived credibility. When looking at how the exposure impacts the individuals' policy views, they find that there is little change.

When looking at whether AI can write persuasive propaganda, Goldstein et al. found that large language models such as GPT-3 can generate propaganda that is nearly as effective as propaganda generated by foreign actors [56]. In their work they used news articles that were part of covert propaganda campaigns and used GPT-3 to generate articles on the same topic. In their work, they found that the propaganda created, both the original and the GPT-3 generated, was highly effective in persuading the respondents.
In their work, they propose that the use of language models can make propaganda campaigns less costly and be on a larger scale with minimal human effort.

4.3.2 Different strategies

In their work, “Truth, lies and automation”, Buchanan et al. tests GPT-3 for different strategies in influence operations on social media [9]. As they describe in their work, there are several aspects to creating disinformation, and they describe how GPT-3 performs in each of these. The strategies and their results are as follows:

Strategy	Description	Performance
Narrative Reiteration	Generating short messages that advance a particular theme.	GPT-3 excels with little human involvement
Narrative Elaboration	Developing a medium-length story that fits within a desired worldview when given only a short prompt	GPT-3 performs well, and fine-tuning leads to consistent performance
Narrative Manipulation	Rewriting news articles from a new perspective, shifting the tone, worldview, and conclusion to match an intended theme	GPT-3 performs reasonably well with little human intervention or oversight, though their study was small
Narrative Seeding	Devising new narratives that could form the basis of conspiracy theories.	GPT-3 easily mimics the writing style of QAnon and could likely do the same for other conspiracy theories.
Narrative Wedging	Targeting members of particular groups, often based on demographic characteristics, with the goal of prompting certain actions or amplifying divisions.	A human-machine team is able to craft credible, targeted messages in minutes. GPT-3 deploys stereotypes and racist language in its writing for this task.
Narrative Persuasion	Changing the view of targets.	A human-machine team is able to devise messages on two international issues – withdrawal from Afghanistan and sanctions on China – that prompt survey respondents to change their positions.

Table 4.1: Truth, lies and automation results [9]

The results from Buchanan et al. show how well a language model can be used for nefarious purposes. In this project the aim is to look at how these models can be used on social media platforms, so some tasks described will not be that important for this project.

For this project, the main area of interest is narrative reiteration, manipulation, and wedging. As seen from the research done on the troll fabric tweets [7], the

strategy employed there was to "play both fields". With tweets made to target members of a particular political belief. This along with strategies of reiterating certain themes and rewriting stories to present Russia in a more positive way.
As this project will only look at short messages (<200 characters), it will not be looked at rewriting stories at a large scale, but rather shorter messages that aim to change the story. For this project, it is also looked at how well these models can perform with minimal amounts of human interaction. Based on their work [9], little human intervention should be needed to gain good results with the use of language models.

4.3.3 Creating personas

In her master thesis, Bonnerud writes about using language models in combination with psychological evaluation models to create a more personalized language generation [22]. Her findings showed that the use of autoregressive models such as GPT-2 and ERNIE work well in creating personalized text generation. Her approach to this was to feed the model personality traits as part of the input to make it write a certain way.
From her research, she identifies the need for automatic evaluation metrics, alongside human evaluation [22]. The results in her thesis show that language models show promising results in generating text for social media users, but that the personality traits are not necessarily that easy to keep intact, or that it is hard for humans to identify these traits in social media texts, although research shows that personality traits can be found in the text [22].
Looking at other research done on the topic, as well as the changes that came with the next generation of language models, i.e., GPT-3, gives indication of an increasing possibility in making changes in the writing style as well as traits and sentiments. Looking at the research done by Dathathri et al. one can see a possibility of gaining a large range of topics and sentiments in text generation, without having to fine-tune large language models for several topics [57]. With the use of an attribute classifier along with a pretrained language model, they were able to generate text on a variety of topics, with different sentiments. Such a tool in the context of text generation in social media could be used to easily personify various accounts, and create a false belief that the opinions spread by these models are an opinion shared by countless others.

4.4 Somulator

FFI, NTNU and the Norwegian cyber Defence has together worked on a simulator for social media, called "Somulator". For the testing with the Norwegian Defense Cyber Academy, the Somulator was used to give the environment more validity.

The Somulator is built up of various open source social media platform made to replicate similar and well—known platforms, such as Twitter, Facebook, Instagram

and YouTube, as well as a platform for posting news articles [58]. Along the open source alternatives for the well—known platforms, there is also an administrative site that can be used to effectively administrate users and upload content for the different platforms.
The Somulator works as an isolated platform, with a username and password, to ensure that the content uploaded cannot be accessed by outsiders. As well as a separation between the administrative sites and the sites for ease of access for the participants.

4.4.1 Architecture

The architecture of the Somulator is a lightweight setup to enable the use of the Somulator without needing access to cloud computing or a large server locally to host the cluster of web servers.
It is built with the use of docker containers [59] to separate the components of each web-server from each other, and to minimize the usage of computing power. With the separation of these web-servers into their own clusters of containers, it is also easier to handle problems with one server without it impacting the others.

As each web-server is split into containers handling their own sets of tasks, such as presenting the front-end or handling reading and writing to the database, it is structured for easily gathering data after an experiment has been complete.

For this thesis the site was used to administer the users and the content for the Mastodon website, which is an open-source software made to compete with Twitter, while the results were fetched with the use of SSH-access and SQL-statements.

Although this thesis only looks at how a language model can perform when generating shorter texts, the Somulator makes it possible to expand on this in further research, with both testing the language models on longer texts with their news site, and to do more research on the social impacts of influence operations in these domains.

Chapter 5

Design and Implementation

This chapter will describe the design of the pipelines created for the text generation. It will describe each part in detail, and the reasoning and results that led to each component being added.

At the start of the project, it was decided to create a pipeline using an English language model, and a pipeline using a Norwegian language model. At the time, there existed only one Norwegian language model, and it was therefore the natural choice for one of them. For the English language model, it was decided to use a language model of similar size and quality. This caused Facebook's OPT-6.7B [13] to be the choice for the second language model.

5.1 English Pipeline

The English pipeline consists of three main parts. The language model, the translator, and a tool to remove grammatical errors. As you can see in the figure below 5.1 the way the production of tweets was approached in this thesis was to push all prompts through both pipelines, both the first pretrained model and the fine-tuned model. The approach for this for each test is described more clearly in chapter 6.

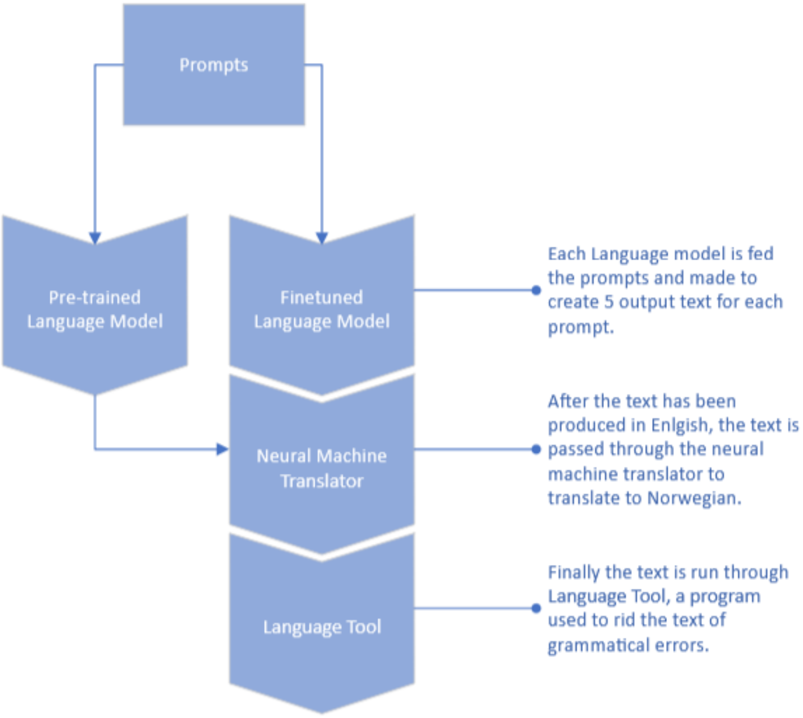

Prompts
Pre-trained Language Model
Finetuned Language Model
Neural Machine Translator
Language Tool
Each Language model is fed the prompts and made to create 5 output text for each prompt.
After the text has been produced in Enlgish, the text is passed through the neural machine translator to translate to Norwegian.
Finally the text is run through Language Tool, a program used to rid the text of grammatical errors.

Figure 5.1: English Pipeline

5.1.1 OPT-6.7B

The model used for generating the English text is the OPT-6.7B model, created by Facebook as an open source alternative to GPT-3. In our pipeline, the pretrained model on the left side is the one trained by Facebook [13]. It has been trained on data from several sources, ranging from books to forum posts. The data the model was trained on is approximately 800GBs of data.

The fine-tuned model was trained on a hyperpartisan news dataset [60]. The idea behind this was to use a dataset that might be able to "play both sides", instead of fine-tuning two models on a dataset for each side. The size of the dataset used was roughly 6 GB in total. The model was fine-tuned for 2 epochs.

5.1.2 OPUS Machine Translator

The neural machine translator used for translating was the OPUS translator made for translating from English to the North Germanic languages, including Norwegian [41].

5.1.3 LanguageTool

After the tweets were translated, they were sent through an application called LanguageTool[61]. This tool was used to remove grammatical errors for each tweet. It was decided to use this tool for its simplicity and possibility to automate the correction. As of the writing of this thesis, it is not freely available to correct the semantics with the use of LanguageTool, but could be used when it becomes available.

5.2 Norwegian Pipeline

The Norwegian pipeline is built in a very similar fashion, with the main difference being that the neural machine translator is removed.

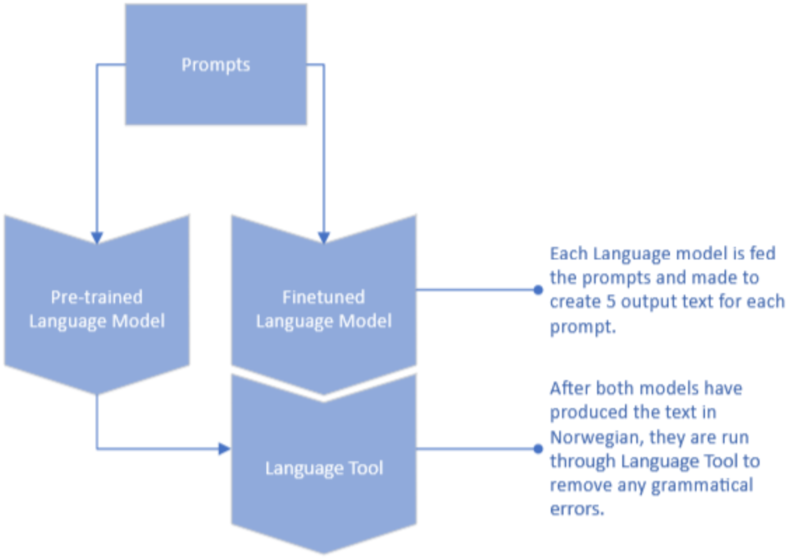
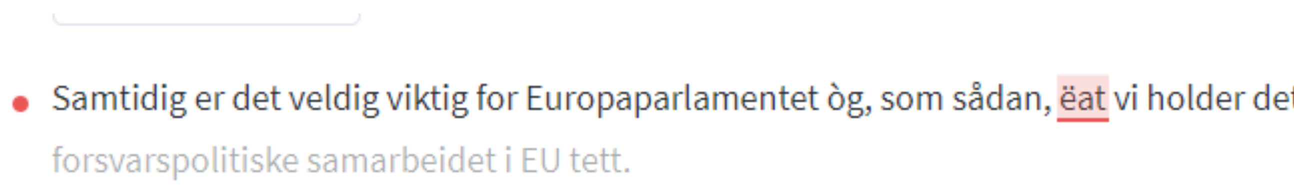

Prompts
Pre-trained Language Model
Finetuned Language Model
Language Tool
Each Language model is fed the prompts and made to create 5 output text for each prompt.
After both models have produced the text in Norwegian, they are run through Language Tool to remove any grammatical errors.

Figure 5.2: Norwegian Pipeline

5.2.1 GPT-J

The Norwegian language model that was used in this pipeline was the GPT-J 6B [14]. This model was fine-tuned by the National library to write in Norwegian,



● Samtidig er det veldig viktig for Europaparlamentet òg, som sådan, ëat vi holder de
forsvarspolitiske samarbeidet i EU tett.

Figure 5.3: LanguageTool example

instead of being trained from scratch on the Norwegian language. In this project, it was then fine-tuned further with a dataset of politicians debating in government [62]. For the generation of tweets, the same prompts was sent through both models, the one created by the national library and the one fine-tuned in this thesis. The model was trained for 3 epochs, but it was decided to go for the model checkpoint after 1 epoch, as the fine-tuning over several epochs caused the model to create gibberish and inconsistent tweets.

5.2.2 LanguageTool

As with the English pipeline, the application LanguageTool was used to remove grammatical errors from the tweets created. For the Norwegian pipeline this was as necessary as with the English pipeline, most likely due to the dataset it was fine-tuned with by the national library. There were some sentences using letters associated with other Nordic countries. The language model also had some spelling mistakes, and would often split composite words. This problem persisted after the fine-tuning of the model.

5.3 Finetuning

With the fine-tuning of the models, a library called Deepspeed [30] was used. As described in the theory chapter, deepspeed is a library that makes it possible to minimize the resources used to train larger models, such as GPT-J. Deepspeed is an engine, which wraps up the model, and handles the data and model parallelism. It also handles the other configurations that one could pass in, such as the precision of the weights (32 or 16 FP). The engine is then passed the script to fine-tune the model, the model to fine-tune, the dataset, etc.

Although the IDUN cluster has large amounts of nodes, GPUs and memory readily available, it was important to implement Deepspeed's techniques as part of the training. Such as the Zero redundancy optimizer. In this project the third stage of the optimizer was applied, which spreads parameters, gradients, and the optimizer state out between all GPUs used. As well as enabling the possibility to offload the optimization memory and computation to the CPU, [30]. This technique is mostly relevant in two separate cases. When training massive models, way beyond the size of the models in this thesis, or when working with a limiting amount of resources. With IDUN this thesis should not fall into either cases, but what was discovered during the training was that when requesting large amounts of resources, the time before the training started would increase to days. To ensure that access to resources would be more instantaneous, it was decided to go with a trade-off. By enabling the third stage of the Zero redundancy optimizer, the process will be slightly slower than when using stage one or two. But it enables the use of all GPU-types available in IDUN, as the models are too big to fit on the smaller ones.

Along with the stage 3 optimization technique, AdamW was implemented as the optimizer. This optimizer, as described in 3, is a memory efficient optimizer, while also being close to achieving the same generalization as SGD [32]. Both models were trained over several epochs, with checkpoints being saved for every step in the epoch. As there are few tests to automate the evaluation of text generators, and even fewer relevant in the Norwegian language, there was a manual evaluation of the models at each checkpoint to find the best one. If any checkpoint seemed to be overfitted, the following checkpoints would also be disregarded. This led to the model that ran for an epoch for the Norwegian model, the GPT-J, and the model created after two epochs for the English model.

Chapter 6

Experiments and Results

This chapter will present the whole process of the experiments done in this thesis, with the goal of presenting the results shown and the steps done to get there. To give researchers a possibility to conduct follow-up tests or use this project as a template for their research. First the setup for each of the experiment, and how the data points were gathered is presented. Finally the results from each experiment is presented.

6.1 Setup

6.1.1 Within-subject design

The pilot tests conducted prior to this thesis were a within-subject design, with the goal of evaluating how two different pipelines performed in writing short text that emulated human writing.
The approach was to first run a test with the pretrained models for each pipeline, both the Norwegian and English. This was done to create a baseline for comparison as well as a possibility to evaluate the two pipelines against each other.
Parallel to the data collection of the first pilot test, the fine-tuning for both models was done.
The survey was generated with the use of Nettskjema [63] and distributed through social media. The survey consisted of 9 tweets, 3 generated by humans, 3 generated with the English pipeline and 3 generated with the Norwegian pipeline. Each participant was asked to evaluate which is created by a human and which is created by a machine. It was also asked for the participants to share their reflection on how they approached the tweets.

The second pilot test was run with the fine-tuned models for each pipeline. The survey for this part was also created with Nettskjema [63] and spread with the use of social media. It consisted of 8 tweets, 4 generated by humans, 2 by the Norwegian pipeline and 2 by the English pipeline. The participants were asked to evaluate their own performance on a scale from 1 to 5, and share their reflection on how they evaluated the tweets.

6.1.2 2x2 factorial within-subject design

For the 2x2 factorial within-subject design, it was of interest to gather data on how the Norwegian language model compared to an English language model, and how the language models perform when writing general texts and texts on concrete problems.

The research design was implemented with a survey created in Nettskjema, and distributed through social media. The survey consisted of 4 groups of tweets, one of a generic nature in English, one of a more concrete nature in English and the same groups in Norwegian. Each category consisted of 4 tweets, 2 created by a model, and 2 created by a human. Other data collected for this test were the participants' age and gender, as well as how certain they were of their answers.

The participants were instructed to mark each tweet they believed was created by a machine for each category. After the participants had gone through each category, they were asked how confident they were in their answers, and also asked to share their reflections on how they approached the tweets.

6.1.3 Embedded research experiment

The embedded research study was done on the 15th of March and took a full day. The first part of the day was used on the tweets that the students were set to evaluate. After this, the students presented their own work with influence operations, before the day was ended with group interviews.

Preparations

Prior to the 15th, the students had a day when they received some lectures on the topic of influence operations, and they were given a task to design their own influence operation to be presented on the 15th.

To prepare for the data collection, it was generated a set of emails and passwords they would use to connect to the social media simulator, along with an ID for us to use to connect their work on the Somulator with their verbal literacy tests, SAM and Judgement of Performance [16, 17].

100 machine generated tweets and 100 human-written tweets was created, and selected 50 tweets from each that would be presented to them on the Somulator. The tweets were selected using a random number tool to ensure that the tweets selected were not impacted by us. After the 100 tweets that would be used for

the testing were selected, they were again given a random order to remove any possible pattern. The amounts of human-written and machine generated for each round of 25 tweets were as follows:

human-written	machine generated
14	11
14	11
14	11
9	16

Table 6.1: Human Machine split for each 25 tweets

Both the human-written and machine generated tweets can be categorized into the 10 sentiments shown in 6.2 below. There were 10 tweets for each category, 5 human-written and 5 machine generated.

pro-Russian	anti-Russian
pro-Ukraine	anti-Ukraine
pro-Norwegian Armed Forces	anti-Norwegian Armed Forces
pro-NATO	anti-NATO
pro-USA	anti-USA

Table 6.2: Tweet categories

In the creation of the tweets, there were put in place some criteria and some rules for how much the tweets could be altered after being created by the language models.

1. The tweet must be between 150 and 200 characters
2. The prompts used for the language models can not be used as part of the tweet
3. Named entities (e.g., President Obama) can be changed when the tweet is describing the present.
4. When the tweet only uses pronouns to describe the person, the person's name can be added to ensure that context is understood.
5. Some tweets can fall into two categories (e.g., pro-Russia and Anti-Ukraine). In these cases, it is the sentiment the tweet concludes with that is chosen as the category.

Experiment in Somulator

Before the students were presented with the tweets in the Somulator a verbal literacy test was conducted. Here they were given instructions on how the test would be conducted. Their task was to write down as many words as they could in a given category, within a minute. They would then be given a new category and repeat the procedure. Names would not count, and words with the same root (e.g., snow and snowman) would only be counted once. The categories they were given were:

1. Words that start with the letter F
2. Words that start with the letter S
3. Words that start with the letter A
4. Words assimilated to animals on the letter F
5. Words assimilated to animals on the letter S

For the words assimilated with animals, names of animals were not allowed. After this, they were instructed to describe how they felt with the SAM scale, and give a judgement of performance on how they believed they would do on the test.

The students were then presented with a test-tweet on the Somulator to show the students how they should respond to the tweets. The way they were told to mark each tweet can be seen in the figures below 6.1 6.2 6.3. To ensure that the students did not forget or mix up which icon was connected to which category, the icon and its corresponding category was written on the whiteboard, available for all to see.

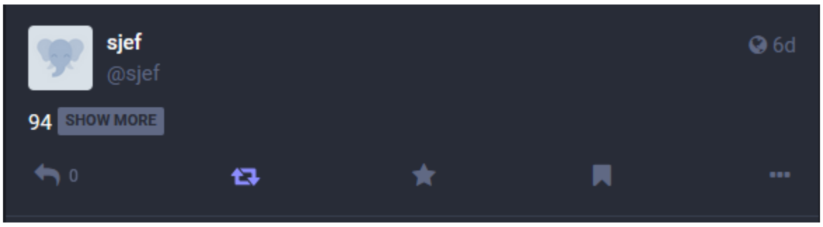


Figure 6.1: Tweet marked as human-written

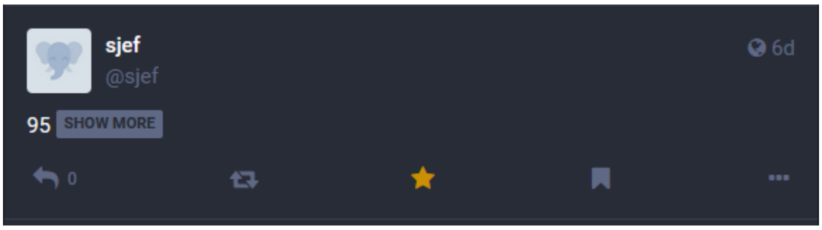


Figure 6.2: Tweet marked as machine generated

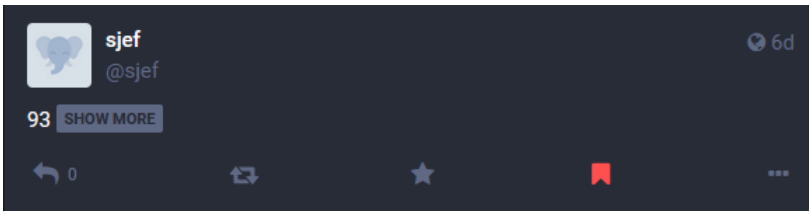


Figure 6.3: Tweet marked as unsure

After this, the real tweets were posted, 5 at a time, every minute. When 25 tweets was reached, they were told to use the SAM scale again and do a judgement of performance for the 25 tweets they had gone through. This procedure was done for each round. The students were given 5 new tweets to evaluate every minute. After 5 minutes, they were told to use the SAM scale and give a judgement of performance for the last 25 tweets they had evaluated. At the end of the testing, they were asked to judge their performance for the full test.

Group interview

The group interview was conducted after both tests were done, and the students had presented their own work. The main focus of the interview was to gain insight into their experience of the experiment, their reflections on the ease of creating content for influence operations with language models. The interview also focused on how the Armed Forces should respond to and mitigate influence operations.

6.2 Results

6.2.1 Within-subject design

Pilot test 1: Pretrained models

Below one can see the results from the first test, with the generic models. Participants were shown 9 texts, one at a time, with the task of identifying if it was written by a human or a machine. In this test, there were 3 text generated by each model, and 3 control texts written by a human. Of these question, 4 of them had more wrong answers than correct, with 2 being human texts, and one from each model.
Participants had an average of **4.84** correct answers, with a standard deviation of **1.74**. Looking at participants' confidence in their answers, one can see that none answered with a **5 - sure**, and most answers were at **3 - neither**, and **2**. None of the participants were able to correctly identify all the texts. Looking at how well participants identified each of the variants, the following can be seen: Human text was identified correctly **51.6**% of the time, the English model **56.9**% of the time, and the Norwegian model **51.6**% of the time.

	Correct	Wrong	Unsure
Text nr 1 (Norwegian)	21	8	2
Text nr 2 (Human)	13	15	3
Text nr 3 (Human)	23	5	3
Text nr 4 (English)	12	12	7
Text nr 5 (Norwegian)	10	16	5
Text nr 6 (Norwegian)	17	13	1
Text nr 7 (Human)	12	16	3
Text nr 8 (English)	20	8	3
Text nr 9 (English)	21	9	1

Table 6.3: Results from first test

1 - unsure	6
2	8
3 - neither	12
4	5
5 - sure	0

Table 6.4: Participants confidence

Pilot test 2: Finetuned models

The results from test number 2 can be seen below. Here the questionnaire was based on 8 texts, 4 written by humans, and 2 by each of the models. With the same task as in the prior test, identifying the ones written by humans and the once written by machines. In this test, 4 of the texts were identified correctly by the majority of the participants, with 2 of them being texts produced by humans, and 2 being from the English model.
Participants had an average of **4** correct answers, with a standard deviation of **1.12**. Out of all participants, none were able to correctly identify all texts. The total results for how well the participants identified each model and the human text is as following: The English model was identified correctly **55.6**% of the time, human text was correctly identified **51.9**% of the time, and the Norwegian model **40.7**% of the time.

	Correct	Wrong	Unsure
Text nr 1 (Human)	7	19	1
Text nr 2 (Norwegian)	11	15	1
Text nr 3 (Norwegian)	11	15	1
Text nr 4 (Human)	9	18	0
Text nr 5 (English)	14	10	3
Text nr 6 (Human)	21	4	2
Text nr 7 (English)	16	9	2
Text nr 8 (Human)	19	8	0

Table 6.5: Results from test 2

6.2.2 2x2 factorial within-subject design

The 2x2 factorial within-subject design was the final test with Nettskjema surveys. Here it was looked at the Norwegian and English model in both general and domain-specific texts. There were 23 participants, **12** female and **11** male. Their distribution in age can be seen in 6.6. As can be seen in the table, there is not a great sample size for each age group, making it hard to do any generalizing data for the age groups.

age	participants
16 - 25	1
26 - 35	13
36 - 45	1
46 - 55	3
56 - 65	3
over 65	2

Table 6.6: Age distribution

	Norwegian	English
Generic	7	6.5
Domain-specific	11	6

Table 6.7: Average correctly identified machine texts

The data in the table 6.7 is calculated by averaging the amount of participants correctly identifying the machine generated texts in each category. There is a significant difference between generic and domain-specific texts in the Norwegian texts. Compared to the English categories, where there is only a small difference.

Categories	text nr	Answered Machine	Answered Human
General Norwegian (cat 1)			
	text nr 1 (machine)	7	16
	text nr 2 (human)	7	16
	text nr 3 (machine)	7	16
	text nr 4 (human)	12	11
Specific Norwegian (cat 2)			
	text nr 1 (machine)	11	12
	text nr 2 (machine)	11	12
	text nr 3 (human)	13	10
	text nr 4 (human)	4	19
general English (cat 3)			
	text nr 1 (machine)	8	15
	text nr 2 (machine)	5	18
	text nr 3 (human)	14	9
	text nr 4 (human)	4	19
specific English (cat 4)			
	text nr 1 (machine)	6	17
	text nr 2 (machine)	6	17
	text nr 3 (human)	4	19
	text nr 4 (human)	9	14

Table 6.8: Total results of 2x2 factorial design

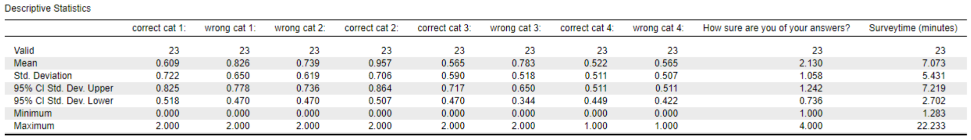
For the table showing the results 6.8 of each text, each text is presented with the correct answer in parenthesis, and the amount of participants answering "machine" and “human” for each text. In this survey, the participants were instructed to only mark texts they believed were generated by a machine. Therefore, the results are evaluated as correct if they correctly mark a machine generated text, and wrong if they incorrectly mark a human-written text as machine. For each category, there are the human-written texts that have the highest number of correct responses for each category. There are some texts written by humans, where many participants believed it was a machine.

Category	correct	wrong
Norwegian (male)	1.727	2
Norwegian (female)	1.417	1.167
Norwegian (Average)	1.565	1.565
English (male)	1.273	1.636
English (female)	0.917	1.083
English (Average)	1.087	1.348
Total (male)	3	3.636
Total (female)	2.333	2.25
Total (Average)	2.652	2.913

Table 6.9: Comparison of rightly identified and incorrectly identified texts between genders

When comparing the average number of correctly marked machine generated texts and incorrectly marking human-written texts, there are some differences between the male and female participants 6.9. Male participants on average identify **0.7** more machine generated tweets than the female participants, but they also incorrectly mark human-written texts **1.386** more than the female participants. When looking at these values for the Norwegian texts and the English texts, the following results can be seen. The male participants on average correctly identify machine generated texts **0.310** more than the female participants, but they incorrectly mark the human-written text **0.833** more than the female participants. For the English text, the same trend can be seen, that the male participants score higher on both correctly identifying machine generated text and incorrectly marking human-written texts. Here however, the results are that the male participants score **0.356** higher on correctly identifying machine generated texts, and **0.553** higher on incorrectly marking human-written texts.

For the table 6.9 comparing the results between gender, there are other interesting factors to look at. On average, females correctly identify the machine generated texts slightly more than they incorrectly mark human-written texts. However, the male participants incorrectly mark human-written texts more than they correctly identify machine generated texts. Their results however are quite low, both in total and for each category. A perfect score would mean 4 correctly identified machine-generated texts for each language, and 8 in total. Meaning the results for both female, and male participants are below 50%, for both languages and in total.

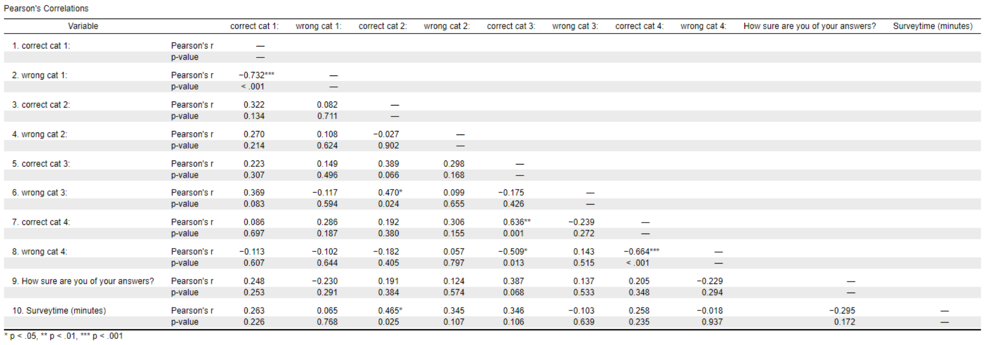


Descriptive Statistics

	correct cat 1:	wrong cat 1:	wrong cat 2:	correct cat 2:	correct cat 3:	wrong cat 3:	correct cat 4:	wrong cat 4:	How sure are you of your answers?	Surveytime (minutes)
Valid	23	23	23	23	23	23	23	23	23	23
Mean	0.609	0.826	0.739	0.957	0.565	0.783	0.522	0.565	2.130	7.073
Std. Deviation	0.722	0.650	0.619	0.706	0.590	0.518	0.511	0.507	1.058	5.431
95% CI Std. Dev. Upper	0.825	0.778	0.736	0.864	0.717	0.650	0.511	0.511	1.242	7.219
95% CI Std. Dev. Lower	0.518	0.470	0.470	0.507	0.470	0.344	0.449	0.422	0.736	2.702
Minimum	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	1.283
Maximum	2.000	2.000	2.000	2.000	2.000	2.000	1.000	1.000	4.000	22.233

Figure 6.4: Descriptive statistics 2x2 factorial design

When looking at the overall results from the 2x2 factorial testing, one can see that there is a noticeable decrease in the mean score of correct answers, going from generic Norwegian to domain-specific Norwegian. However, the mean score of wrong answers rises slightly.
The English categories, on the other hand, stay relatively equal between the categories, with the main changes being in the mean score of wrong answers. From the figure 6.4 it can be seen that there is no one that had 2 correct or 2 wrong answers in the category domain-specific English.



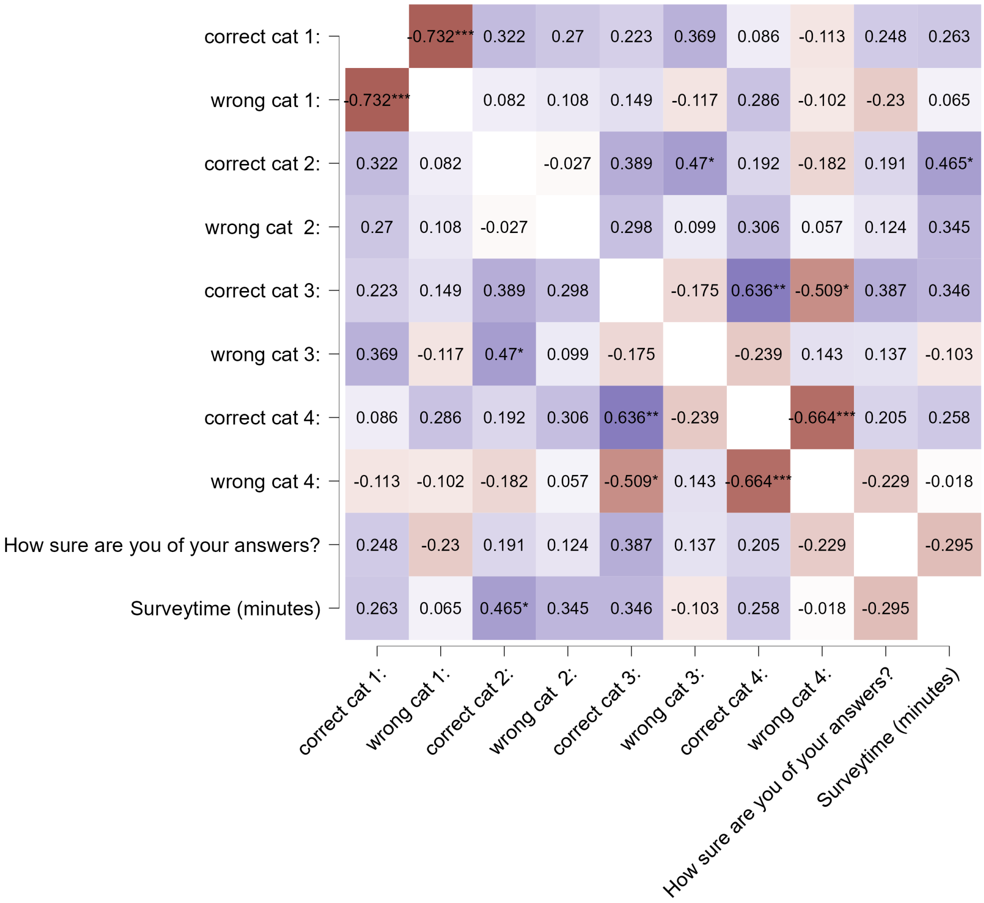
Pearson's Correlations

Variable		correct cat 1:	wrong cat 1:	correct cat 2:	wrong cat 2:	correct cat 3:	wrong cat 3:	correct cat 4:	wrong cat 4:	How sure are you of your answers?	Surveytime (minutes)
1. correct cat 1:	Pearson's r	—									
	p-value	—									
2. wrong cat 1:	Pearson's r	−0.732***	—								
	p-value	< .001	—								
3. correct cat 2:	Pearson's r	0.322	0.082	—							
	p-value	0.134	0.711	—							
4. wrong cat 2:	Pearson's r	0.270	0.108	−0.027	—						
	p-value	0.214	0.624	0.902	—						
5. correct cat 3:	Pearson's r	0.223	0.149	0.389	0.298	—					
	p-value	0.307	0.496	0.066	0.168	—					
6. wrong cat 3:	Pearson's r	0.369	−0.117	0.470*	0.099	−0.175	—				
	p-value	0.083	0.594	0.024	0.655	0.426	—				
7. correct cat 4:	Pearson's r	0.086	0.286	0.192	0.306	0.636**	−0.239	—			
	p-value	0.697	0.187	0.380	0.155	0.001	0.272	—			
8. wrong cat 4:	Pearson's r	−0.113	−0.102	−0.182	0.057	−0.509*	0.143	−0.664***	—		
	p-value	0.607	0.644	0.405	0.797	0.013	0.515	< .001	—		
9. How sure are you of your answers?	Pearson's r	0.248	−0.230	0.191	0.124	0.387	0.137	0.205	−0.229	—	
	p-value	0.253	0.291	0.384	0.574	0.068	0.533	0.348	0.294	—	
10. Surveytime (minutes)	Pearson's r	0.263	0.065	0.465*	0.345	0.346	−0.103	0.258	−0.018	−0.295	—
	p-value	0.226	0.768	0.025	0.107	0.106	0.639	0.235	0.937	0.172	—

* p < .05, ** p < .01, *** p < .001

Figure 6.5: Correlation matrix 2x2 factorial design

When looking at the correlation matrix 6.5 and heatmap 6.6 for the 2x2 factorial research, there are three strong correlations that should be highlighted. First, there is a strong correlation between correct answers in category 3 and correct answers in category 4. This pattern can not be seen between category 1 and category 2.
There is also a strong negative correlation between correct answers and wrong answers in category 2. Indicating that participants mainly answer correct or wrong in this category. The same negative correlation can be seen between correct and wrong in category 4.



	correct cat 1:	wrong cat 1:	correct cat 2:	wrong cat 2:	correct cat 3:	wrong cat 3:	correct cat 4:	wrong cat 4:	How sure are you of your answers?	Surveytime (minutes)
correct cat 1:		-0.732***	0.322	0.27	0.223	0.369	0.086	-0.113	0.248	0.263
wrong cat 1:	-0.732***		0.082	0.108	0.149	-0.117	0.286	-0.102	-0.23	0.065
correct cat 2:	0.322	0.082		-0.027	0.389	0.47*	0.192	-0.182	0.191	0.465*
wrong cat 2:	0.27	0.108	-0.027		0.298	0.099	0.306	0.057	0.124	0.345
correct cat 3:	0.223	0.149	0.389	0.298		-0.175	0.636**	-0.509*	0.387	0.346
wrong cat 3:	0.369	-0.117	0.47*	0.099	-0.175		-0.239	0.143	0.137	-0.103
correct cat 4:	0.086	0.286	0.192	0.306	0.636**	-0.239		-0.664***	0.205	0.258
wrong cat 4:	-0.113	-0.102	-0.182	0.057	-0.509*	0.143	-0.664***		-0.229	-0.018
How sure are you of your answers?	0.248	-0.23	0.191	0.124	0.387	0.137	0.205	-0.229		-0.295
Surveytime (minutes)	0.263	0.065	0.465*	0.345	0.346	-0.103	0.258	-0.018	-0.295	

Figure 6.6: 2x2 factorial design Heatmap

6.2.3 Embedded Research Design

The participants in the case study consisted of 35 students, 26 male and 9 female. Of these 35 students, only 10 participated in both the testing with the Somulator tweets and the group-interview.

Experiment in Somulator

On average, the students identified **51.3%** of the tweets correctly. The lowest accuracy was at **33%**, while the highest was at **67%**, with the CI Mean being [48.279, 54.349]. In table 6.10 you can see the variance in the average between the genders, as well as between each set of 25 tweets.

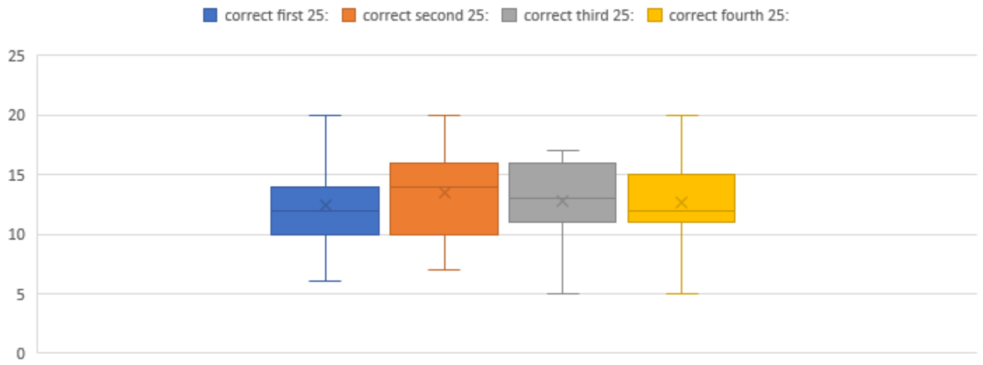


Figure 6.7: Box and whisker plot for correct answers per 25 tweets

	male avg	female avg	total avg
first 25 tweets	11.96	13.88	12.45
second 25 tweets	13.26	14	13.45
third 25 tweets	12.23	14.22	12.74
fourth 25 tweets	12.76	12.33	12.65
total	50.23	54.44	51.31

Table 6.10: Average correct identified tweets

As only marking tweets as human or only machine would give an accuracy of 50%, it was also looked at the precision of the students' evaluation of tweets. In the table 6.11 the students' precision when identifying machine generated tweets is presented. As there was a varying amount of machine-tweets in each round of tweets, there are different baselines for the different groups. Similar to the score of 50% that could be used as a baseline when looking at the total score, the baseline when looking at precision is as follows:

$$Precision\,Baseline = \frac{Amount\,of\;machine\,tweets\,in\,round}{Total\,amount\,of\;tweets\,in\,round}$$

Each round has 25 tweets, while the amount of machine generated tweets vary. For the first three rounds, there were 11 machine generated tweets in each, giving us a baseline of 0.44. Meaning that only answering machine would give us a precision of 0.44. For the last round, the amount of machine generated tweets were 16, giving us a baseline of 0.64. For the total, a precision of 0.5 is the baseline. When looking over the data, it was identified that some students had answered machine on every tweet in a specific category.

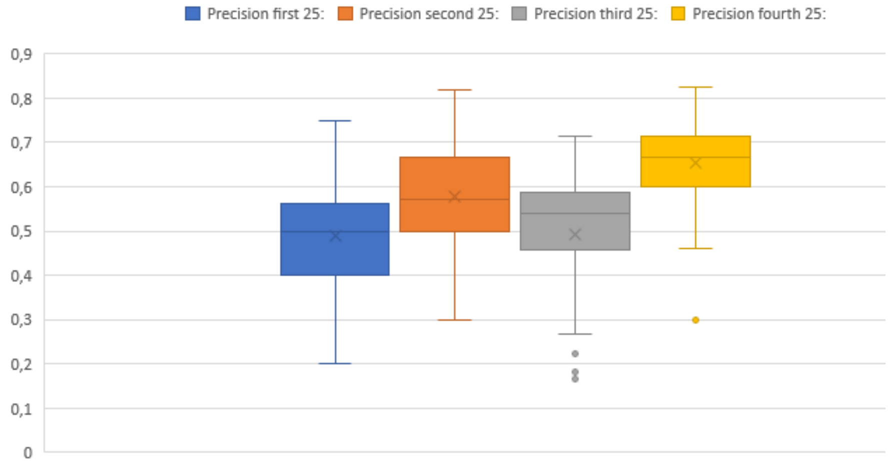


Figure 6.8: Box and whisker plot for precision per 25 tweets

	male avg	female avg	total avg
first 25 tweets	0.4795	0.5189	0.4896
second 25 tweets	0.5810	0.5714	0.5785
third 25 tweets	0.4848	0.5186	0.4935
fourth 25 tweets	0.6492	0.6638	0.6530
total	0.5536	0.5701	0.5579

Table 6.11: Average precision identifying machine generated tweets

Looking at the tables describing the average correct identified tweets, 6.10 and the precision 6.11 there are slightly better results among the female participants. There is however not a large enough sample size to draw any conclusions to the larger population. The average correctly identified tweets stay roughly at 50% for each round, with the biggest variance being the second round, which sits at 53.8%. When looking at the table showing the precision, there is a bigger variance in the second round, however, with a precision of **0.5785**.

Table 6.12: Paired Samples T-Test

Measure 1		Measure 2	t	df	p
correct first 25:	-	correct second 25:	−1.311	34	0.199
correct first 25	-	correct fourth 25:	−0.255	34	0.800
correct second 25:	-	correct fourth 25:	1.169	34	0.251

Note. Student's t-test.

It was also performed a paired samples T-test on how well students perform between the different rounds, to see whether fatigue has any impact on their

ability to perform. The third round was removed from the comparison, due to its low scores when testing for normal distribution.
As can be seen from 6.12 there are no strong indications of fatigue having an impact between the rounds.

Category	Accuracy	Precision
Pro-Ukraine	0.5857	0.6342
Anti-Ukraine	0.5057	0.4939
Pro-Russian	0.5571	0.6091
Anti-Russian	0.4886	0.5325
Pro-USA	0.4229	0.4608
Anti-USA	0.4714	0.5130
Pro-NATO	0.5857	0.6484
Anti-NATO	0.4686	0.5106
Pro-Armed Forces	0.4971	0.5474
Anti-Armed Forces	0.5457	0.5973

Table 6.13: Accuracy and Precision in categories

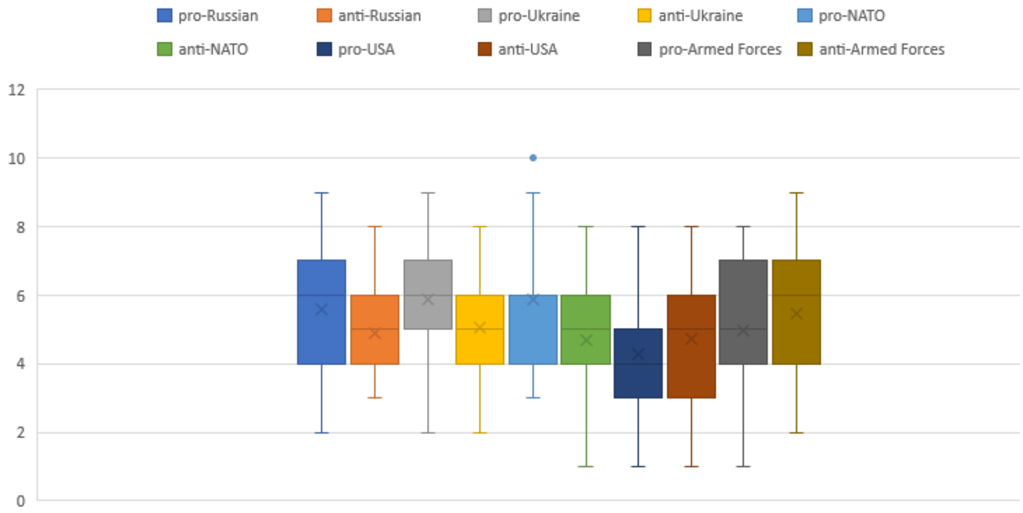


Figure 6.9: Box and whisker plot of accuracy on the different categories

Looking at the accuracy and precision in each category in table 6.13 one can see that most stay close to 0.5. This is the baseline described earlier. For most of the categories, the students score better at identifying the pro-category compared to the counterpart. The exceptions are for the categories on the USA and armed forces. Here, the students score lower on the pro, compared to the anti-category. The categories that can be reckoned as slight outliers are; Pro-Ukraine, Pro-NATO and Pro-USA. When looking closer at these tweets, no explanatory factors was found in the data. To see if there were any traits that could have an impact on the outcome, some analysis of the individual tweets was conducted.

The length of each tweet was analyzed, and how many correct answers there were for each tweet, to see if there was a correlation there. The result was **0.00507**, showing no sign of correlation between the two variables. When looking at the correlation between the students' evaluation of how well they did in the prior round, with how well they actually did, there was a correlation of **-0.06024**. When looking at the same correlation, but factoring in how sure they were of their own evaluation, the correlation still remained low, at **-0.0516**. Showing no signs of correlation between their own evaluation and performance.
When looking at the students' confidence before a round, and their own results, a slightly higher correlation can be seen, at **0.1376**. This correlation is still not high enough however, to claim any correlation between the two variables.

In their research on how an IT-background has an impact on participants' metacognitive accuracy, confidence and overestimation in ability to identify deep fakes, Sütterlin et al. used what they called the Overconfident scale (OCS)[64]. In their research, they found that the results from the OCS was a good indicator of participants in need of more followup training. As the data collected in this experiment was the same data as Sütterling et al. on the participants' self evaluation, it was decided to calculate the same variable in this thesis. To see if their findings could also be applied to this research The variable, OCS, is calculated with the following formula:

$$OCS = \frac{(\frac{Pre-CIA*100}{11}) + 1}{\% of\ correct\ ratings + 1}$$

Where Pre-CIA (confidence in abilities) is the participants' confidence in how well they will perform in the following round. That is divided by 11, as that is the degree of freedom the participants have when answering that question. The score calculated from the formula will describe how the participants' self-evaluation aligns with their accomplishment. A score below 1 means the participants underestimate their own performance, while a score above 1 means the participants overestimate their own abilities. The OCS was calculated for each round, as well as plotting the plots for how the students' confidence changed throughout. This can be seen in the figures 6.10 6.11. In these box plots, there are some interesting differences between the male and female participants. The female participants have a tighter spread, with a mean closer to 1. The male participants start the first round with a mean OCS score of **1.33**, before dropping closer to 1.

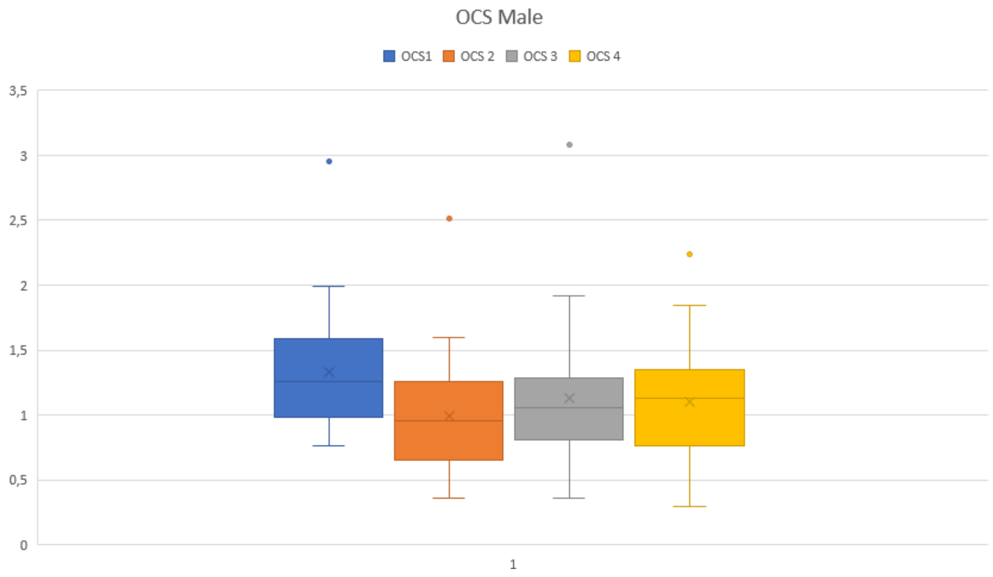


Figure 6.10: Over Confident Scale male participants

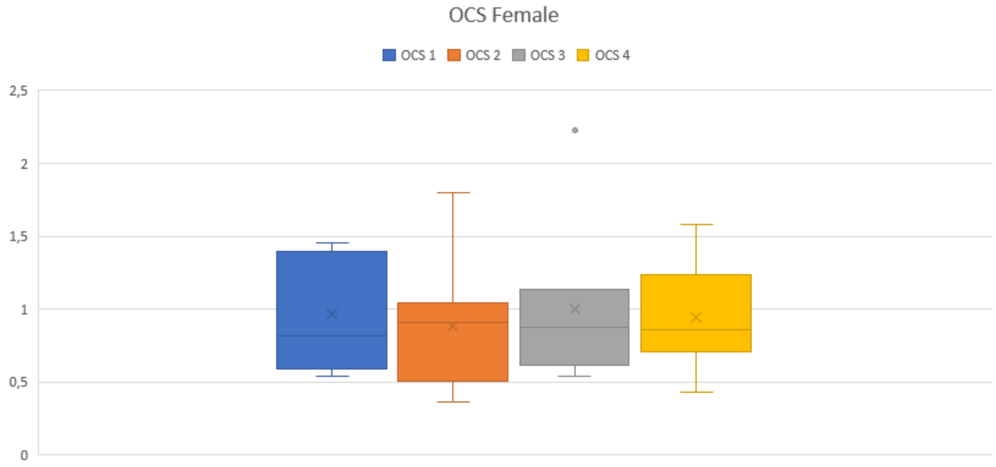
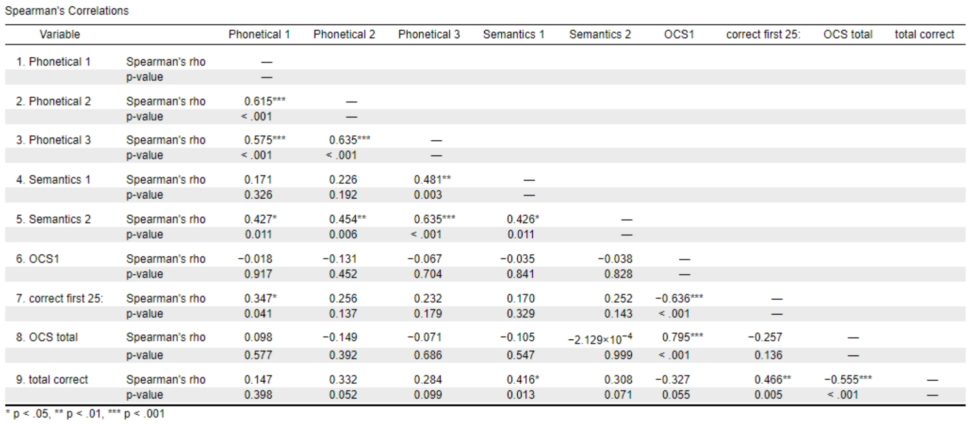


Figure 6.11: Over Confident Scale female participants

Looking at the correlation matrix and heatmap 6.13 6.12 for the variables from the experiment, there is a moderate negative correlation between the OCS from the first round and the % of correct answers in the first round. There is also a moderate negative correlation between the total OCS and the total % of correct answers. For the pre-tests, there was a low correlation between the semantics and the total score.



Spearman's Correlations

Variable		Phonetical 1	Phonetical 2	Phonetical 3	Semantics 1	Semantics 2	OCS1	correct first 25:	OCS total	total correct
1. Phonetical 1	Spearman's rho	—								
	p-value	—								
2. Phonetical 2	Spearman's rho	0.615***	—							
	p-value	< .001	—							
3. Phonetical 3	Spearman's rho	0.575***	0.635***	—						
	p-value	< .001	< .001	—						
4. Semantics 1	Spearman's rho	0.171	0.226	0.481**	—					
	p-value	0.326	0.192	0.003	—					
5. Semantics 2	Spearman's rho	0.427*	0.454**	0.635***	0.426*	—				
	p-value	0.011	0.006	< .001	0.011	—				
6. OCS1	Spearman's rho	−0.018	−0.131	−0.067	−0.035	−0.038	—			
	p-value	0.917	0.452	0.704	0.841	0.828	—			
7. correct first 25:	Spearman's rho	0.347*	0.256	0.232	0.170	0.252	−0.636***	—		
	p-value	0.041	0.137	0.179	0.329	0.143	< .001	—		
8. OCS total	Spearman's rho	0.098	−0.149	−0.071	−0.105	-2.129×10^{-4}	0.795***	−0.257	—	
	p-value	0.577	0.392	0.686	0.547	0.999	< .001	0.136	—	
9. total correct	Spearman's rho	0.147	0.332	0.284	0.416*	0.308	−0.327	0.466**	−0.555***	—
	p-value	0.398	0.052	0.099	0.013	0.071	0.055	0.005	< .001	—

* p < .05, ** p < .01, *** p < .001

Figure 6.12: Correlation matrix

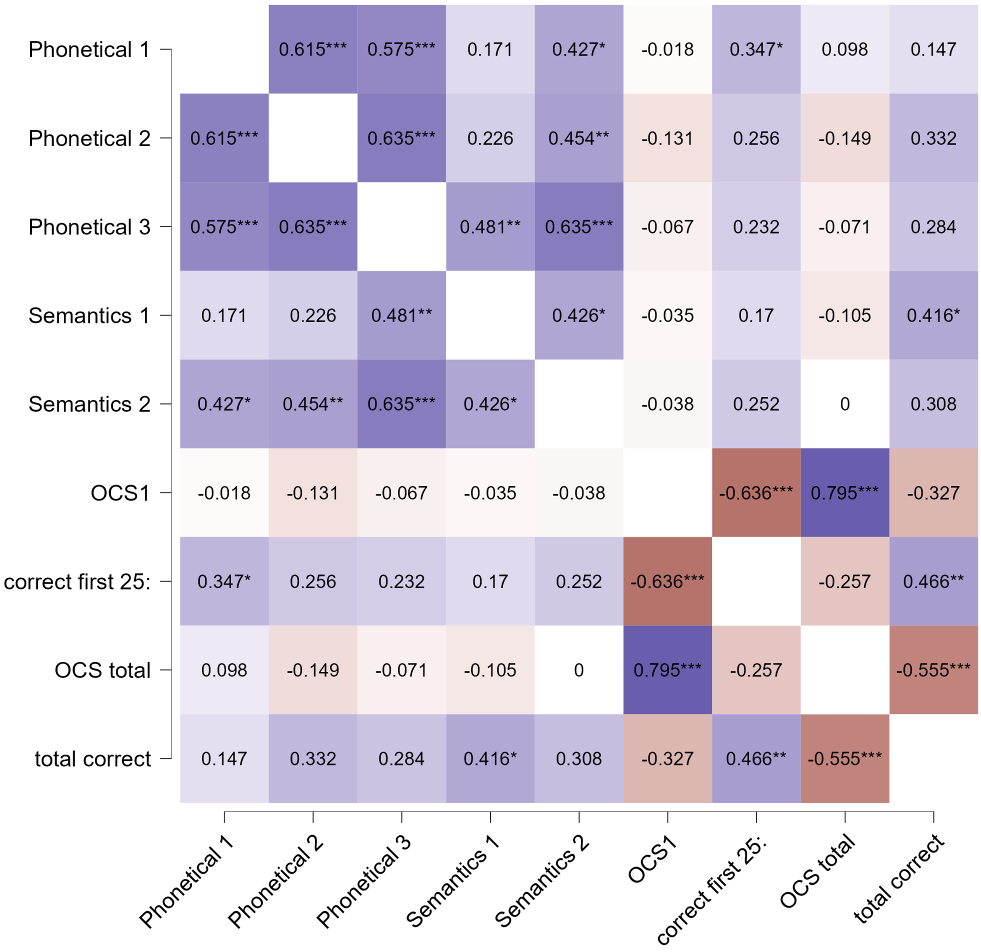


Figure 6.13: Heatmap

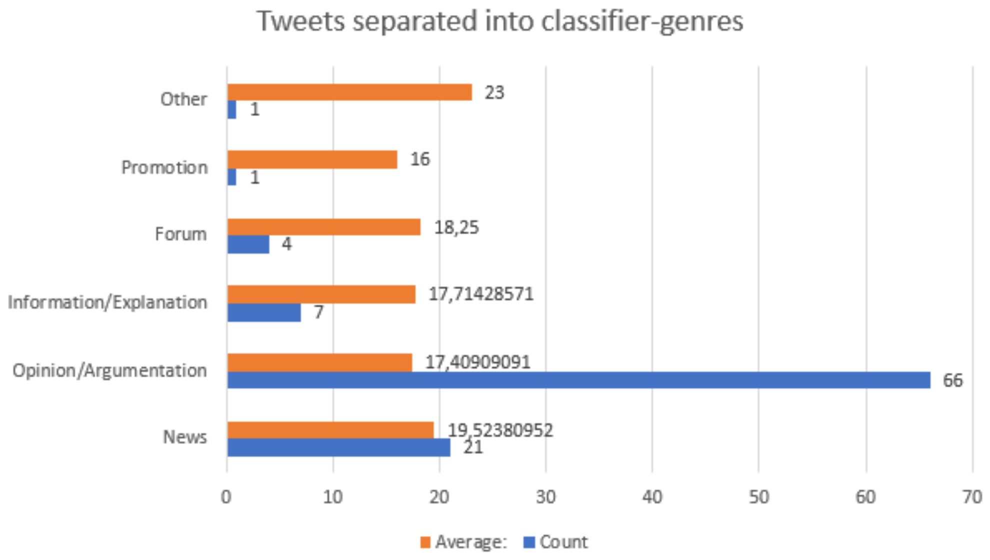
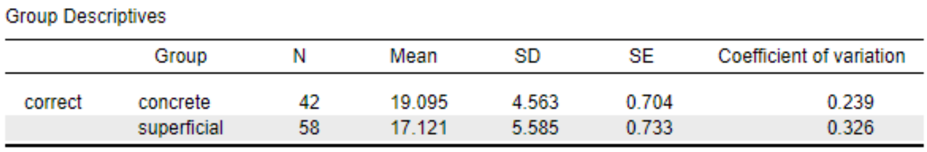


Figure 6.14: Count of each category and the average correct answers

Each tweet was also analyzed with a classifier used to categorize texts into different categories [65]. The produced results from the classifier can be seen in figure 6.14. The main categories were "Opinion/Argumentation" with 66 of the 100 tweets, and “News” with 21 of the tweets. The 4 other categories were "Information/Explanation" with 7 tweets, “Forum” with 4 of the tweets and “Promotion” and “Other” with one tweet each. The average correctly identified tweets divided by classified category is roughly the same, with “News” being the only slight outlier of any relevance. “Other” and “Promotion” have a slightly bigger deviation from the rest, but with only one tweet, it has no statistical significance.

Another way it was tried to group the tweets were into the generic nature of the text. Tweets were split into two groups: generic and concrete. The deciding factor for whether a tweet was put in the generic group or the concrete group was if the tweet could actually be connected to a concrete event or action. If the tweet could be critic or praise of any kind of event or action, it was marked as generic.

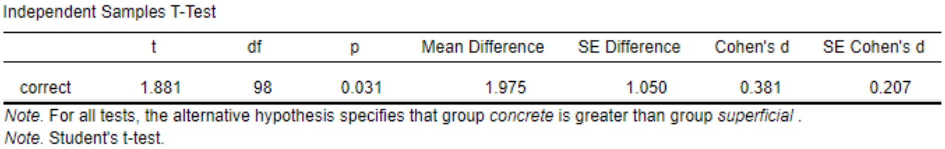


Group Descriptives

	Group	N	Mean	SD	SE	Coefficient of variation
correct	concrete	42	19.095	4.563	0.704	0.239
	superficial	58	17.121	5.585	0.733	0.326

Figure 6.15: Descriptive statistics of generic and concrete tweets

As can be seen in the figure 6.15, there were a slightly bigger group of generic tweets than there were of concrete tweets. There is also a big difference between the confidence interval of the generic tweets and the concrete tweets. But when doing a one-tailed independent t-test, testing for the hypothesis that concrete texts are more easily detected than the generic tweets, the following score 6.16 is the result. Showing a significant difference between the groups.



Independent Samples T-Test

	t	df	p	Mean Difference	SE Difference	Cohen's d	SE Cohen's d
correct	1.881	98	0.031	1.975	1.050	0.381	0.207

Note. For all tests, the alternative hypothesis specifies that group *concrete* is greater than group *superficial* .
Note. Student's t-test.

Figure 6.16: t-test of generic and concrete tweets

When looking specifically at machine-tweets, there is a higher p-value at 0.154. But still the same trend, indicating that the generic tweets are harder to spot.

6.2.4 Group Interview

In the group-interview there were initially 16 who had announced their participation, but due to dropout, the group participating consisted of 10 students. Prior to the experiment, they were given lectures on how influence operations work and instructed to make their own influence operations. They were divided into 8 groups of the same size, with different tasks regarding the influence operation. They were given a fictive country based on the OCCASUS-scenario often used in NATO exercises [66], and either the goal of reinforcing the population's thrust in their government or to sow distrust. The students were given free reins to solve the task as they themselves saw fit. For this reason, there were several solutions, such as using false profiles imitating political figures, or creating large amounts of accounts to give users a false sense of what the majority supports. There were also numerous approaches when it came to the use of different tools. Some decided to not use any sorts of tools, while others used meme generators as well as language models to aid in their operation. Their work from this assignment was also part of the group interview, and gave some interesting reflections from the students when it comes to the importance of educating people on the topic. As well as their perception of the threats posed by influence operations, assisted with language models.

Influence operations in general On the topic of influence operations in general, their reflections revolved around how concealed these operations can be. Some interview subjects had made their influence operations not only by publishing opinions and information to sway people, but also enhanced these opinions with other techniques. Examples included overloading the platform so that other opinions are hard to find, and giving users of the platform an impression that the opinions are held by the majority of users.
During this topic, they also brought up their own mindset prior to the testing, and

how it changed during the testing. Prior to their exposure to the Somulator tweets, several of the interview subjects described being very sure of their own abilities in detecting the machine generated tweets. During one of the rounds, one of the interview subjects realized they had a tendency to believe the tweets that went against his own views were written by a machine, and the opposite for those that aligned with his views. It was also commented by the informants that some texts had a structure that reminded them more of an English sentence structure than a Norwegian one.

When asked about what should be done to handle these challenges and what tools and capabilities people should have to detect influence operations, their answers were primarily aimed at more information and knowledge. To gain knowledge of how it is used, experiencing influence operations in safe environments, and to make this knowledge available to as many as possible. There were also suggestions of running campaigns to remind people of thinking critically when engaging on social media. There were also suggested to research the use of algorithms to isolate attempts of influence operations. The interview subjects appeared to all agree that the best solution would be to counter attempts of influence operation with as much real information as possible.

AI enhanced influence operations When discussing how machine learning models could be used and in what way they enhanced influence operations, the interview subjects had used several tools and models they could reflect upon. One of the main aspects they drew attention to was how easily accessible the tools were, and how little technical knowledge they needed to use the models. Their use of language models was primarily with ChatGPT [67], a language model released with beta-access in November 2022.

One of the informants described a feeling of starting to understand how the language model they had been using would respond, but felt that this was hard to apply in these tests, as he did not know which model had been used in this setup. In general, they believed that to be able to separate between influence operations enhanced or run by AI, from influence operations run by humans. Instead, they believed that the biggest problem surrounding AI enhanced influence operations were how it gave even a single individual a possibility to run an operation at a large scale, similar to what Goldstein et al. describes in their paper [56].

Chapter 7

Discussion

In this chapter, the results from this research is discussed comparisons drawn to research found in our literature review to answer our research questions. This chapter is also used to draw some light towards additional findings made in this research. Finally, the limitations of the research and suggested improvements is presented.

Similar to other low-resource languages, there has been very little work done on language models in the Norwegian language. Prior to this thesis, there was only one language model trained to generate Norwegian texts [40]. This model was fine-tuned on an already existing English model, instead of creating one from scratch. For other NLP-tasks, such as translating, there have been used a model from another language as an initial point for transfer learning. For translation, this can be done with good results when doing transfer learning from a multilingual model to a fine-tuned model for translating a specific language pair [68]. For language models that are not multilingual from the outset, no research has been done on how it will impact the final results. When conducting the experiment, there were some interesting comments during the group interview. Some Informants commented that several of the tweets had a sentence structure that aligned more with the English sentences. As we did not do a cross-check after the group interview to see which of the tweets the students pointed to, one can only speculate in whether they noticed a real pattern, and if it was tied to the model or if it was a pattern that could be seen both in the human written texts and machine written texts.

7.1 Research Question 1

Is the perceived authenticity of the text being generated by the language model affected by any of the following factors:

The goal of this research question was to get a broader understanding of how language models perform in the Norwegian language. With sub-questions focused on how different factors impacted the perceived authenticity. The answer to this research question forms the basis for further research on factors that makes it hard for a human evaluator to detect that a text is generated by a language model.

7.1.1 Research question 1.1

The use of different datasets in the creation of the language model

As described in chapter 4 the use of datasets that better match the task can give an improved performance [9]. The first pilot tests, aimed at answering this research question. The results from these tests show clear indications that the perceived authenticity was strengthened with the use of datasets to fine-tune the models toward concrete topics. The results from the following tests, both the 2x2 factorial design and the Somulator experiment, showed the same results. As the amount of correctly identified texts remained low for the generated text from the fine-tuned models.

When looking at the results from the first pilot test in the within-subject design, the average score of correctly identified texts is 4.84, with a quite large deviation, of 1.74. It is interesting to note that the Norwegian and English model are both identified correctly as robots 2/3 of the time. While the texts written by humans are identified correctly, only 1/3 of the time. For one of the texts in the first pilot survey a major problem emerged, as the human texts were chosen from the Norwegian Colossus Corpus (NCC) [69]. Since the texts used to populate the human side of the questionnaire were picked arbitrarily, it had the possibility of being a text that made little sense. The participants have shown that the language models are hard to detect when generating shorter texts similar to social media content. From this study, it can also be shown that text written by humans goes under scrutiny when participants are told to find the machine generated texts.

Results from the second test 6.5 shows a decreased ability to identify machine generated texts from both models. However, there was a bigger improvement for the Norwegian model. The average here was at 4, with a deviation of 1.12. In this test, both the texts generated by the English model were identified by the majority as written by a machine. Neither of the texts written by the Norwegian model were identified by the majority of participants, as only 40% answered correctly on each of these questions. It should be specified that two of the texts written by humans were wrongly identified. With one of them being incorrectly identified by 70% of the participants, and the other only 33% of participants were able to correctly attribute to the human category.
An explanation for this might be the domain that the texts are written about. Gender equality in the Armed forces, and the climate change debate, are topics that have had a large focus in the media the past years. From research done by

Movarec et al. [5] Users on social media platforms have a higher chance to believe news that aligns with their own views. Future work should gather data of participants political views prior to testing. With the goal to better understand how much of an impact this has on the participants' answers. This also enables the possibility to look at how much the content made from the language models can alter a persons initial view, as Kreps, McCain and Brundage investigates in their work [55].

Looking at the total results from both the first and second pilot test, one can see that the amount of times the human-written text is identified correctly stays roughly the same, only differing with **0.2**%. The English model has a slight drop, from **56.9**% to **55.5**%. The biggest change came with the Norwegian model, with a drop from **51.6**% to **40.7**%.
This indicates that the datasets have a positive impact on its perceived authenticity, albeit stronger for the Norwegian model than for the English model. Due to the small amount of texts tested in the within-subject design, it should be done further testing. To better understand how the pipelines compare and which areas they differ in.

The testing from the pilot study only looked at how each text on its own was evaluated. Without any larger context added to the text, or comparison between texts. The texts were also presented in an unnatural environment, through the survey. As there was no time limit, the participants could use as much time as they wanted to evaluate each text. In the case of twitter, Counts & Fisher found that the mean time used on a single tweet was 2.92 seconds [15]. Therefore, the results from the within-subject design can show us that these models perform well in generating Norwegian texts in shorter format. And that the finetuning of the models with different datasets make the model more capable of writing more concrete texts on the topics connected to the dataset.

7.1.2 Research question 1.2

The domain the model generates text about

Throughout the different test designs, the texts generated have related to different domains. From the experiment, the most comprehensive testing was done, of how well participants performed with different domains. Here the texts were divided into different topics; Ukraine, Russia, NATO, USA, and the Norwegian Armed Forces. These topics were split into positive and negative sentiments. It was also conducted some tests in the 2x2 factorial design test, where texts of different topics was presented in different contexts. One where all were texts about the same topic, and one where the texts were of different topics.

For the 2x2 factorial design testing, it was of interest to do a comparison of models of similar size to get a better understanding of how they differ. Prior research that has looked at how well the language models perform in English, has either looked at models from a previous generation, like GPT-2, or models far bigger than in this thesis. To get a better comparison, it was decided to use the models fine-tuned in this thesis, which are roughly the same size. As the data available to compare against was of the largest language models, it was not possible to do a good comparison of how this language model performed.
These texts were also presented in a survey, but with a different presentation. The texts were presented in groups to better match how a user on social media could read messages that were either part of their regular feed or grouped together due to a search term. The participants were then asked to mark all texts in each group they believed were written by machines. There were no time limits in this test either.
On average, each participant scored below 50% correct in each category. However, for the generic Norwegian texts, there were stronger results than there were for the domain-specific texts. When looking at the differences between the English categories, there were minuscule differences between the groups. There was also a strong negative correlation between correct and wrong answers in category 1 (generic Norwegian) and the same negative correlation could be seen in category 4 (domain-specific English). For category 3 (generic English) there was a weak correlation, while there was none for category 2 (domain-specific Norwegian).

The domain-specific English texts revolved around the American election in 2016 and the domain-specific Norwegian texts revolved around the war between Ukraine and Russia. For that reason, it is hard to do find an explanation for why there is such a big difference between the Norwegian and English scores in the domain-specific category. A possible explanation is that the participants' knowledge of the topic has an impact on how well the models score, and should therefore be further researched. Another possible explanation is that the datasets used for the Norwegian model was older, and created before the conflicting war. While the English model was trained on a dataset that was largely built on news stories around the American election in 2016.
In the pilot study, as well as the Somulator experiment, each participant was required to answer either "human", "machine" or "unsure" for each text. If we look away from the "unsure"-option, only answering "human" or "machine" gives a participant a 50% score. This can then be seen as the baseline.
In the 2x2 factorial test the texts were presented in groups, and the participants were tasked with marking each text in the group they believed to be generated by a machine, the results differed from the pilot test. If participants believed no texts were generated by a machine, they would not mark any text. This gave us an average score for each category well below 50% for all categories, except the domain-specific Norwegian. Indicating that these models could go undetected to a larger degree when there is not a known presence of a language model.

The results from the Somulator experiment showed that the topic had an impact on the perceived authenticity. The students scored the lowest on the topics regarding the USA. Here, both the pro and anti tweets had an average score below 50%. With the pro-tweets being at 42.6% and the anti-tweets being at 47.1%.
When looking at reasons for the low score on these tweets compared to the other tweets, there are some possible explanations. Most of the tweets about the USA focused on things from the past, with focus on the last decades. When looking at the tweets about Russia, Ukraine and the Norwegian Armed Forces, one can see that a lot of the tweets are centered around the present, with conflicts and debates that are currently in focus. The tweets about NATO also had a focus on past actions as well, but with an average score higher than those about the USA. Some tweets about the USA also focused on polarizing topics such as the current and former presidency. A possible explanation for the bad scores in these categories could be that their political view has an impact on their answers [55].
Another possible explanation is the familiarity that the students have with the different domains. Based on their background, they are kept up to date on the war between Ukraine and Russia. As well as educated and informed on NATO and the Armed Forces. There is a possibility that their knowledge of the USA is at a lower level than their knowledge of the other domains. This should be investigated further with tests evaluating participants' knowledge, prior to exposing them to machine generated tweets on the topics. This, to understand how a person's knowledge of a domain can have an impact on their accuracy.

The topic with the biggest difference between positive and negative sentiment is NATO, with a difference of 1.17. Here the students, on average, correctly identified 5.86 tweets for the pro-NATO tweets, and only 4.69 for the anti-NATO tweets. For the rest of the topics, there was a roughly 0.6 difference between the average score of the pro and anti tweets of each topic.
A good explanation for this difference in the data was not found. It is therefore proposed to test more on how negative and positive sentiments affect the score within a topic. To see if there can be identified any possible explanation for these differences.

7.1.3 Research question 1.3

The generic nature of the text being created

With this research question, the goal was to understand how the perceived authenticity was impacted by how concrete the text was. Concrete in the context of this thesis is based on whether the text can be pinned to a concrete event or action. In the results, it was done a one-tailed independent t-test with the hypothesis that concrete tweets would be detected more often than the generic tweets. The one-tail test, instead of the two-tail test, presents a danger of missing the real connection between the groups, since the relationship is only examined in one direction.

For that reason, it was first tested with a two-tailed t-test which showed strong signs of a difference between the groups, but not enough to mark a significant difference. The one-tailed relationship showed that the generic tweets were identified significantly less than the concrete ones. When testing concretely on machine tweets, the results were slightly higher. This indicates that the difference between generic and concrete tweets are bigger when the tweets are human-written.

To better understand how the detail-level of the tweets impact, more detailed rules should be set in place to separate generic tweets from concrete tweets. It should also be tested with models trained on different datasets, and where the tweets better match the dataset.
The reason for this is that the dataset used to fine-tune the model stretches far back in time, and does not have concrete data related to the war in Ukraine and Russia. Nor on current news regarding NATO and the Armed Forces. The use of a different dataset for fine-tuning, or using the current model on different topics, could show different results regarding the generic nature of the tweets.

7.1.4 Research question 1 conclusion

To conclude, on **RQ 1** one can see that the use of different datasets has an impact on the perceived authenticity of the texts being generated. By narrowing the scope of the model by fine-tuning them, the perceived authenticity has increased. Both when fine-tuning the model on a subset of the data the model was pre-trained with. And when fine-tuning the model on a completely new dataset.
The domain of which the model generates text about, and whether the model's text were presented along with text regarding the same domain also had an impact on the perceived authenticity. With the model having a decreased perceived authenticity when being presented with domain-specific text. The experiment also shows that there were some differences in how well the students performed in identifying the machine-generated tweets in each domain. As described earlier, there can be several factors that cause these differences. When looking closer at any possible internal explanation, none was found. Research shows that social media users are more inclined to believe fake news if they support their own views [5]. How this impacts a persons' ability to evaluate whether the text is written by a human or machine is not known, however. Some participants from NDCA spoke of their bias, making them evaluate texts with an opposing world view being marked as generated by a machine. These reflections are in line with the findings from Kreps', McCain's and Brundage's research [55]

There is clear from our research that language models can perform well for shorter texts in low-resource languages, and that with the use of curated datasets, these models can be perceived as authentic in numerous areas. The 2x2 factorial research design, showed signs of participants being more reserved in marking tweets as machine-written. Another interesting consequence, shown in the tests conducted in this thesis, is how the known presence of a language model makes the participants scrutinize the human-written texts.

7.2 Research Question 2

How can a Norwegian language model be used in the context of a cyber-social range for training purposes?

The second research question asked in this thesis is related to the Somulator used in the experiment, and the work done by FFI in their project C-SPI [70]. The reason for this question was to find out to what degree a language model can be used for nefarious intentions. And if a language model can be used in any way to make exercises in the Somulator less costly both in time and resources.

7.2.1 Research question 2.1

Can the language model be used as a tool for creating influence operations?

There are several ways a language model could be used as a tool for training purposes, and in the creation of an influence operation. As described by Buchanan et al. [9] there are various ways an influence operation could be enabled by a language model. Ranging from simple techniques such as narrative reiteration, where the language model only reiterates already existing narratives, to more complex techniques like narrative seeding. Here the narrative is created by the language model. In this thesis, the work has been focused on narrative reiteration, where it has shown good results.

From the different results in this thesis, various ways this language model can be used as a tool for training purposes has presented itself. With variation in both how much human intervention has been allowed for the different tests, and the approach to how participants would evaluate texts, the results show different aspects that should be taken into consideration when using the language model in an exercise.

The model performs the best when the participants are not instructed to evaluate each text individually, but select the texts from a group that they believe are machine-written. This, regardless of whether the texts were presented in a generic group or domain-specific group.
These results indicate that the model is capable of filling the Somulator with general content, as well as content that is part of an influence operation. The model has also shown its capability to create content on several topics and with both positive and negative sentiments. This creates the opportunity to create exercises for various aspects without having to change the language model.
As described in the related works, chapter 4, there are techniques that would

only add a small extra layer, but make the model more capable of writing texts with different political viewpoints [57]. For future work with the language model produced in this thesis, it should be looked at combining it with the technique proposed by Dathathri et al. in their research[57]. This technique involves combining language models with a classifier that guides the language model in the text generation without any further training of the language model [57].

7.2.2 Research question 2.2

To what extent can the model operate independently?

In this thesis, it has been the goal to minimize the human intervention on the output made by the machines. With the aim of getting better data to answer this research question. As the tests have been very different, it has also been varied how much human intervention has been allowed for each test. For the pilot study and the 2x2 factorial design testing, the only human intervention was in writing the prompts for the models. If the models used the wrong named entities, it was not corrected in these tests. The reason for this was to gain a better understanding of how well the model performed without any human intervention.
When prompts were given to the language model, it was instructed to generate 5 output texts for each prompt. Out of these, those that were deemed best were chosen. There were also several prompts that did not make texts that were deemed good enough to be used. These findings indicate that there is a need for some human quality assurance in creating the prompts and extracting the best outputs. A possible solution that should be looked at is implementing tools to the pipeline that can aid in this prosess.

When conducting the Somulator experiment, the only allowed human intervention was to change the named entities in the output from the model. Here, similar to the other tests, there were also created various outputs for each prompt, and there were used numerous prompts that did not generate good enough results. Causing a need for manual work to extract the 100 machine generated tweets.
One of the main issues that surfaced during the creation of the tweets used for the experiment is that the language model is not trained on a dataset that is recent enough. Making it hard to create texts on current trends and topics that has happened after the models' dataset was collected. Although, it seems to perform well when writing generic texts, and texts that do not focus on named entities. The results from this thesis indicate that it can operate independently to some degree when given prompts, as long as the text is generic or does not focus on named entities. When named entities are used, there has to be some human intervention involved to ensure that the model is not spotted due to its mistakes, i.e. when naming the president in Ukraine or the minister of defense in Norway.

It might be of interest when using the Somulator to use the model in different ways than has been done in this thesis. When using the Somulator in the exper-

iment, every tweet was published with the same user, to minimize uncontrolled variables. The analysis from Linvill and Warren on tweets from the IRA [7], show how operators in the IRA manage several fake users. These fake users would be on both sides of the political spectrum in the USA. In future work with the Somulator it should be looked at the possibility of using the language model in the same way. Using it to produce content that can be posted through different fake users.
In this research, there are no indications of this being possible without human intervention. Both the creation of prompts, and sorting the output to the correct fake users would need human intervention. Future work should look at the possibility of using the technique described by Dathathri et al. [57] and the use of other users' tweets as prompts to solve this problem. To see if the language model can follow the expected behavior of a user and to better understand the limitations of a language model when operating independently.

7.2.3 Research question 2.3

What concrete tasks can the language model fulfill?

In this research, it has been looked at how a Norwegian language model performs in writing short texts, like social media posts. And how such a language model compares to a language model in a high-resource language such as English. Finally, it has been looked closer at how such a language model can perform tasks related to training purposes in a cyber-social range.
The results from the first sets of testing in the pilot study shows clear indications of the perceived authenticity being strengthened with the use of datasets to fine-tune the models toward concrete topics. These results are again reinforced during the 2x2 factorial design, where the texts produced by the fine-tuned models were perceived as more authentic.

The results from the Somulator experiment along with the data from the 2x2 factorial design and the pilot study shows clear signs of language models being capable of writing short texts perceived as authentic by evaluators.
In the context of the Somulator and for training purposes, the language model is fully capable of handling the task of generating larger quantities of data to fill the Somulator. It has also shown that it is capable of generating content that mimics opinions from both sides of debated issues.
From the results in this thesis, it has also become clear that the model is dependent on human intervention at several steps of the process to ensure that the content matches the context. The way the model and pipeline is currently structured, it is not capable of performing tasks individually, without any human evaluation before the output is used.

7.2.4 Research question 2 conclusion

In the context of a cyber-social range, such as the Somulator, and for training purposes, the language model has shown qualities that can make it a useful tool for automating the workload. Both in generating content for influence operations in training contexts and content to fill the Somulator with innocuous content.
The results of this thesis also indicate that these language models are a lot more available and easily trainable with less machine resources due to tools such as Deepspeed [30]. With this increased availability, it can be expected that there is also an increased possibility of individuals using these language models for nefarious intentions. In their work, Goldstein et al. finds that the use of language models can cause as trustworthy propaganda as human-made with very little human effort[56], further proving the point.
In this thesis, there has been a focus on using only the bare minimum of necessary resources, to ensure that the jobs on IDUN do not end up stuck in queues for longer periods of time. As well as to assess whether the language model can operate in the context of the Somulator, as there is no guarantee of access to a super computer to run the model. It can be stated that the model is capable of handling some tasks, such as narrative reiteration. And that it is capable of producing generic content for the Somulator. Since it is capable of running inference on only 1 GPU without a massive impact on run time. This makes it suitable in cases where the Somulator won't have access to large amounts of computing power.

7.3 Additional findings

During the experiment, we gathered more data than was used to answer the research questions of this thesis. From our testing with the students we gathered several metrics that we could use to measure how the students' performance went, their feelings during the testing as well as pre-tests on how their verbal literacy was. When looking at these data points, we have found some data that could explain a persons' performance to a degree. But we have yet to find any data that is capable of predicting a persons' performance to a larger degree.
When looking at the pre-tests, we could see a slight correlation between the final score and one of the semantic tests. When we set up a multiple regression model to see how well these variables predict the final score, we got a correlation of **0.450**, with an adjusted R^2 score of **0.125**. These findings show us that a participants' verbal literacy can be used to a certain degree to predict the participants' final score, but it is not a strong predictor of the final score. In the Armed Forces, there are several standardized tests in use to evaluate soldiers in their recruitment period. With the use of the results from these tests, it could be possible to do more research, to see if there are any test-variables that better reflects a participants' performance.

There was no strong correlation between the students' confidence or self-efficacy and their results. Neither when looking at their confidence before nor after the round. When looking at the OCS however, there is a strong correlation. The comparison of confidence and performance was done prior to this thesis in a study by Sütterlin et al. [64]. Here they found the OCS to be a useful tool when identifying participants in need of more training and follow up. From our results, we share the same findings, with a high correlation between the OCS and the final score for the students. The metric is easily attained and gives good indications of those that have a very skewed understanding of how well they will perform compared to how well they actually do. Similarly to them, we see the same strong correlation in this thesis, indicating that this comparison is a good tool to identify those most in need of more education and training to be better quipped with these language models.

There is also another point that has been made more clear from these tests, when participants have been presented with these texts and instructed to identify which are generated by machine and which are written by humans. Not only are they having a hard time correctly identifying the machine-generated, but it also has a big impact on the perceived authenticity of the human-written texts. This should be looked into further to investigate how big an impact this has on people when using social media platforms. As the known presence of a language model reduces the credibility of genuine texts.
When describing how AI can aid in defeating fake news, Anne and George Cybenko mention how social networks, microblogs, and news outlets cater to sub-populations with their idiosyncratic opinions. Which in turn can lead to these sub-populations believing their opinion is the opinion of the majority [8]. From the results of this thesis it indicates that these language models could be used to cause the same outcome, by automating the production of viewpoints that might be idiosyncratic, but are perceived as the opinion of the majority.

The final point that should be made from the data collected from the experiment is an interesting finding when looking at the amount of "machine"-answers given by the students in each domain (pro-Russia, anti-Russia, etc.). As there were 35 participants and 10 tweets in each category, the total amount of answers per category is 350. Meaning, a balanced result would lead to 175 answers for "human" and 175 for "machine", if the participants possibility of answering "unsure" is ignored. For the Ukrainian and NATO tweets, it was close to balance. When looking at the Russian tweets there was a big difference however, with the pro-Russian tweets being marked as machine far more and an equal shift the other way with the anti-Russian tweets. When looking at the tweets about the USA, both the pro and anti tweets were marked a lot more as machine. The opposite can be seen with the tweets about the Armed Forces, where the tweets are marked a lot less as machine than 175 per category.
It cannot be concluded from this study what the reason for these differences are,

but it gives us some indications that there might be similar personal biases when evaluating these tweets as is the case when evaluating what is fake news and not in social media [5]. When looking at the research conducted by Kreps, McCain and Brundage [55] it is made clear that a persons partisanship influences the perceived credibility. Pointing in the same direction as our data.

7.4 Limitations

This section will highlight some limitations and weaknesses of the research conducted in this thesis.
The within-subject design tests conducted at the earliest stage of this thesis concluded that the Norwegian pipeline was the one that would be used for the next phases of the thesis. Due to the time frame of the thesis, it was not possible to do more testing before making a decision, but the conclusion might have been hasted. The results from the 2x2 factorial design test shows that the English categories score better than the Norwegian categories. Given these results, it gives indications that the English pipeline should have been tested out further with different translators instead, as this might have given a better result.

During the group interview that was conducted after the experiment, it was mentioned by some students that several of the tweets had a sentence structure that reminded them more of a typical English sentence structure than a Norwegian sentence structure. It would be of great interest to see whether the English sentence structure was actually a pattern from the language model that they spotted, or if this pattern was recurring in both the human-written tweets and the machine-generated tweets. Due to a late realization of this observation, it was not possible to perform a review of the tweets the students believed had an English pattern.

The experiment also introduced the timing of students, an aspect that was not present in the former testing. In the experiment, the students were given a full minute per 5 tweets, or 12 seconds per tweet if evenly distributed. When looking at how much time an average user uses on a single tweet, the results were 2.92 seconds[15]. Although the time given to the students is far greater than the average time a user spends on a single tweet, there are still problems with this method. In a natural environment, however, the time spent is not due to a restraint, nor is the task given one that requires analytical thinking. The introduction of the time limit is a possible stress factor in this study that can have impacted the students' results negatively.

Another weakness of the experiment has been the granularity when analyzing the tweets. The texts were placed either in the positive or negative category of a domain. And texts that could be part of several categories was strictly placed in

only one. Having more sentiment-categories, such as neutral, and very negative/-positive would give more detailed results. But would also require more work and a larger amount of tweets in total.
Another aspect was that tweets were only divided into generic or concrete. Analyzing the complexity of words and sentences would greatly improve the data gathered in the Somulator experiment.

Despite these limitations, this thesis provides valuable insights into the use of language models in a low-resource language. Both to better understand its threats when used in influence operations, and how they can be used for training purposes in the Somulator.

Chapter 8

Conclusion and further work

8.1 Conclusion

In this thesis it has been looked at how language models can perform in the context of generating short texts, in the Norwegian language, similar to the length of tweets. It has also been the goal of the thesis to understand how these language models can be used for training purposes in an isolated social network. The central questions for this research were as follows:

1. Which factors impact the perceived authenticity of the text being generated by the language model?
2. How can a Norwegian language model be used in the context of a cyber-social range for training purposes?

To answer these questions, it has been developed two different pipelines for generating text. One implementing a Norwegian transformer-based language model and one implementing an English transformer-based language model, which then translates the text with the use of a neural machine translator. The research design that has been used was a within-subject design, a 2x2 factorial within-subject design and an experiment with collection of pre-test data as well as a group interview afterward with the participants.

To evaluate the models, it was first conducted a within-subject design where participants were presented with texts written by each model to identify which performed the best and would be focused on in the following tests. These pilot tests were conducted twice, once with the pretrained models and once after the models were fine-tuned with datasets aimed at writing text on these polarizing topics. Here the Norwegian language model performed the best, and was the one used for the 2x2 factorial design and case study that was conducted later in the thesis. The results from the tests also show a decrease in participants' accuracy when the models are fine-tuned.

From the 2x2 factorial design, the results indicated that the texts produced by these language models are more capable of going undetected than what the other tests indicated. When the participants were not instructed to give a conclusive answer to each text, it resulted in fewer correctly identified machine-written texts.
The results from this testing also showed that the English texts were slightly harder for participants to identify in both the generic and domain-specific category. For the Norwegian categories, the generic performed better than the domain-specific. The possible explanation made for the lower scoring in the domain-specific category is that the datasets used to fine-tune the Norwegian language model does not contain data on the topic that was used. If these language models are to be used in a Somulator there should be some degree of human intervention to ensure that the text produced actually matches the current state of affairs.

Both the within-subject design tests and the 2x2 factorial within-subject design tests were distributed through the use of surveys, giving the participants unlimited time to answer the survey, and an unnatural environment for the presentation of the texts. To better see how the texts were evaluated in a more natural environment, the case study was conducted with the use of an isolated social network dubbed the Somulator.
The first two research designs were also conducted without any tweaking on the output from the pipeline. In the case study, it was decided to allow changes to named entities to better understand how these language models can be used for training purposes.

In the case study, participants were given a more natural environment as they were presented with the text in the form of social media posts on a website instead of being presented with texts in a survey. Here, there was also a time limit in regard to how long time a participant could use on evaluating each tweet.
Although the time given was longer than the average time used on social media [15], the time limit can have caused some stress reactions from the participants that would not be there naturally in a regular setting.

Supported from the prior testing, the data from the case study shows that the domain the model generates text about has an impact on the perceived authenticity. It is also made clear that the generic nature of the texts produced has a big impact on the text's perceived authenticity, as the concrete texts are correctly identified at a higher rate than the generic ones. This, even though these tweets were corrected so that the named entities matched the current state of affairs.
From the data gathered from the pre-tests, it can be drawn a correlation between the participants' semantic performance and their end result. It is however, not a strong predictor of the end results.

As mentioned above, the 2x2 factorial design showed how the model lacks in data related to current events. When these named entities are corrected afterward,

these models are capable of completing several tasks related to training purposes. They are fully capable of producing large amounts of text that can then easily be filtered through by a human to gather the best results. They have also shown capabilities of creating perceived authentic texts on a range of topics, which makes them useful for many scenarios.
From the work throughout this thesis, it has become clear that the model is not capable of performing on the same level without any human intervention.

8.2 Contributions

First and foremost, this thesis has developed and designed two pipelines for the purpose of creating content for an isolated social network used for training purposes. The models are fine-tuned to create content centering around polarizing topics of the last couple of years.
The results from this thesis have made an initial work in understanding how different factors impact the perceived authenticity of the texts produced. It has also shown that language models in a low-resource language like Norwegian can perform at levels close to what we see in language models in high-resource languages, when the texts are short.

8.3 Further work

The data gathered in this thesis is mostly focused on getting a broad understanding of how these language models perform and how they can be used for training purposes. For that reason, there are several aspects of the research done in this thesis that should be looked further into.

Future studies should also focus on understanding participants reasoning when concluding on whether a tweet is from a language model or not. As mentioned in the discussion, this is an area that was not dived deeper into in this thesis, but the results from the group interview show a plethora of social and psychological aspects to analyze.

The texts produced by the language models have never been presented in a natural setting. Although the case study aimed at creating a more natural environment, it did not quite get there. For further work, it should be looked into how the language models perform when the social network consists of natural users.

In the same focus area, it should also be looked at how the language model performs in aiding the training event organizers. During this thesis, when looking at whether the models can be used as a tool for the event organizers, it has also been part of testing to see how the models perform. The results from the 2x2 factorial design testing indicates that the models are harder to identify when the participants are not instructed to evaluate each tweet individually. It would be of great interest to better understand how well they can aid in the tasks related to these exercises, as there might be similar results seen here, when the models are used without informed participants.

Bibliography

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, *Attention is all you need*, 2017. DOI: 10.48550/ARXIV.1706.03762. [Online]. Available: https://arxiv.org/abs/1706.03762.

[2] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever and D. Amodei, 'Language models are few-shot learners,' *CoRR*, vol. abs/2005.14165, 2020. arXiv: 2005.14165. [Online]. Available: https://arxiv.org/abs/2005.14165.

[3] O. J. Aasen, 'Large models and small languages – using ml to generate social media content for training purposes,' Department of Computer Science, NTNU – Norwegian University of Science and Technology, Project report in IMT4205, Jun. 2019.

[4] C. Paul and M. Matthews, *The Russian "Firehose of Falsehood" Propaganda Model: Why It Might Work and Options to Counter It*. Santa Monica, CA: RAND Corporation, 2016. DOI: 10.7249/PE198.

[5] P. L. Moravec, R. K. Minas and A. R. Dennis, 'Fake news on social media: People believe what they want to believe when it makes no sense at all.,' *MIS Quarterly*, vol. 43, no. 4, Dec. 2019.

[6] A. Bergh, *Social network centric warfare – understanding influence operations in social media*, https://www.ffi.no/publikasjoner/arkiv/social-network-centric-warfare-understanding-influence-operations-in-social-media, Oct. 2019.

[7] D. L. Linvill and P. L. Warren, 'Troll factories: Manufacturing specialized disinformation on twitter,' eng, *Political communication*, vol. 37, no. 4, pp. 447–467, 2020, ISSN: 1058-4609.

[8] A. K. Cybenko and G. Cybenko, 'Ai and fake news,' *IEEE intelligent systems*, vol. 33, no. 5, Sep. 2018.

[9] B. Buchanan, A. Lohn, M. Musser and K. Sedova, 'Truth, lies, and automation,' Center for Security and Emerging Technology, Tech. Rep., 2021.

[10] L. Fröhling and A. Zubiaga, 'Feature-based detection of automated language models: Tackling gpt-2, gpt-3 and grover,' eng, *PeerJ. Computer science*, vol. 7, e443–e443, 2021, ISSN: 2376-5992.

[11] A. ALDayel and W. Magdy, 'Stance detection on social media: State of the art and trends,' eng, *Information processing & management*, vol. 58, no. 4, p. 102 597, 2021, ISSN: 0306-4573.

[12] P. D. Leedy and J. E. Ormrod, *Practical Research Planning and Design*. Pearson, 2015.

[13] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang and L. Zettlemoyer, *Opt: Open pre-trained transformer language models*, 2022. DOI: 10.48550/ARXIV.2205.01068. [Online]. Available: https://arxiv.org/abs/2205.01068.

[14] P. E. Kummervold, J. De la Rosa, F. Wetjen and S. A. Brygfjeld, 'Operationalizing a national digital library: The case for a norwegian transformer model,' in *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, 2021, pp. 20–29. [Online]. Available: https://aclanthology.org/2021.nodalida-main.3/.

[15] S. Counts and K. Fisher, 'Taking it all in? visual attention in microblog consumption,' The AAAI Press, Jan. 2011. [Online]. Available: https://www.microsoft.com/en-us/research/publication/taking-visual-attention-microblog-consumption/.

[16] J. Egeland, N. I. Landrø, E. Tjemsland and K. Walbækken, 'Norwegian norms and factor-structure of phonemic and semantic word list generation,' *The Clinical Neuropsychologist*, vol. 20, no. 4, pp. 716–728, 2006, PMID: 16980257. DOI: 10.1080/13854040500351008. eprint: https://doi.org/10.1080/13854040500351008. [Online]. Available: https://doi.org/10.1080/13854040500351008.

[17] T.-M. Bynion and M. Feldner, 'Self-assessment manikin,' in Jan. 2017, pp. 1–3. DOI: 10.1007/978-3-319-28099-8_77-1.

[18] R. Dale, 'Natural language generation: The commercial state of the art in 2020,' eng, *Natural language engineering*, vol. 26, no. 4, pp. 481–487, 2020, ISSN: 1351-3249.

[19] A. Gatt and E. Krahmer, 'Survey of the state of the art in natural language generation: Core tasks, applications and evaluation,' eng, *The Journal of artificial intelligence research*, vol. 61, pp. 65–170, 2018, ISSN: 1076-9757.

[20] R. DALE, 'Nlp in a post-truth world,' eng, *Natural language engineering*, vol. 23, no. 2, pp. 319–324, 2017, ISSN: 1351-3249.

[21] A. Sherstinsky, 'Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network,' *Physica D: Nonlinear Phenomena*, vol. 404, p. 132 306, Mar. 2020. DOI: 10.1016/j.physd.2019.132306. [Online]. Available: https://doi.org/10.1016%2Fj.physd.2019.132306.

[22] K. Bonnerud, 'Write like me: Personalized natural language generation using transformers,' eng, M.S. thesis, Norwegian University of Science and Technology, 2021. [Online]. Available: https://hdl.handle.net/11250/2835483.

[23] A. Radford and K. Narasimhan, 'Improving language understanding by generative pre-training,' 2018.

[24] R. Srinivasan and A. Chander, 'Biases in ai systems: A survey for practitioners,' *Queue*, vol. 19, no. 2, pp. 45–64, Apr. 2021, ISSN: 1542-7730. DOI: 10.1145/3466132.3466134. [Online]. Available: https://doi.org/10.1145/3466132.3466134.

[25] B. Banitz, 'Machine translation: A critical look at the performance of rule-based and statistical machine translation,' eng ; por, *Cadernos de tradução*, vol. 40, no. 1, pp. 54–71, 2020, ISSN: 1414-526X.

[26] D. Shterionov, R. Superbo, P. Nagle, L. Casanellas, T. O'Dowd and A. Way, 'Human versus automatic quality evaluation of nmt and pbsmt,' eng, *Machine translation*, vol. 32, no. 3, pp. 217–235, 2018, ISSN: 0922-6567.

[27] M. Själander, M. Jahre, G. Tufte and N. Reissmann, *EPIC: An energy-efficient, high-performance GPGPU computing research infrastructure*, 2019. arXiv: 1912.05848 [cs.DC].

[28] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh and H. Wu, *Mixed precision training*, 2017. DOI: 10.48550/ARXIV.1710.03740. [Online]. Available: https://arxiv.org/abs/1710.03740.

[29] S. Rajbhandari, J. Rasley, O. Ruwase and Y. He, *Zero: Memory optimizations toward training trillion parameter models*, ArXiv, May 2020. [Online]. Available: https://www.microsoft.com/en-us/research/publication/zero-memory-optimizations-toward-training-trillion-parameter-models/.

[30] S. Rajbhandari, J. Rasley, O. Ruwase and Y. He, *Zero: Memory optimizations toward training trillion parameter models*, 2019. DOI: 10.48550/ARXIV.1910.02054. [Online]. Available: https://arxiv.org/abs/1910.02054.

[31] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, 2014. DOI: 10.48550/ARXIV.1412.6980. [Online]. Available: https://arxiv.org/abs/1412.6980.

[32] I. Loshchilov and F. Hutter, *Decoupled weight decay regularization*, 2017. DOI: 10.48550/ARXIV.1711.05101. [Online]. Available: https://arxiv.org/abs/1711.05101.

[33] A. B. Yoo, M. A. Jette and M. Grondona, 'Slurm: Simple linux utility for resource management,' in *Job Scheduling Strategies for Parallel Processing*, D. Feitelson, L. Rudolph and U. Schwiegelshohn, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 44–60, ISBN: 978-3-540-39727-4.

[34] M. Sun, H. Wang, M. Pasquine and I. Abdelfattah Abdelhameed, 'Machine translation in low-resource languages by an adversarial neural network,' eng, 2021, ISSN: 2076-3417. [Online]. Available: https://hdl.handle.net/11250/2831132.

[35] T. Ranasinghe and M. Zampieri, 'Multilingual offensive language identification for low-resource languages,' eng, *ACM transactions on Asian and low-resource language information processing*, vol. 22, no. 1, pp. 1–13, 2023, ISSN: 2375-4699.

[36] A. Kumar, C. Esposito and D. A. Karras, 'Introduction to special issue on misinformation, fake news and rumor detection in low-resource languages,' eng, *ACM transactions on Asian and low-resource language information processing*, vol. 22, no. 1, pp. 1–3, 2023, ISSN: 2375-4699.

[37] B. Wang and A. Komatsuzaki, *GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model*, https://github.com/kingoflolz/mesh-transformer-jax, May 2021.

[38] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo and Y. Iwasawa, *Large language models are zero-shot reasoners*, 2022. DOI: 10.48550/ARXIV.2205.11916. [Online]. Available: https://arxiv.org/abs/2205.11916.

[39] Y. Fu, H. Peng, L. Ou, A. Sabharwal and T. Khot, *Specializing smaller language models towards multi-step reasoning*, 2023. DOI: 10.48550/ARXIV.2301.12726. [Online]. Available: https://arxiv.org/abs/2301.12726.

[40] P. E. Kummervold, J. De la Rosa, F. Wetjen and S. A. Brygfjeld, 'Operationalizing a national digital library: The case for a norwegian transformer model,' in *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, 2021, pp. 20–29. [Online]. Available: https://aclanthology.org/2021.nodalida-main.3/.

[41] J. Tiedemann and S. Thottingal, 'OPUS-MT — Building open translation services for the World,' in *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal, 2020.

[42] R. R. Mackey, *Information warfare*, Accessed: 2022-04-26, 2014. [Online]. Available: https://www.oxfordbibliographies.com/view/document/obo-9780199791279/obo-9780199791279-0024.xml.

[43] B. Tarman and M. F. Yigit, 'The impact of social media on globalization, democratization and participative citizenship,' *Journal of Social Science Education*, vol. 12, no. 1, Jun. 2012.

[44] S. Talwar, A. Dhir, D. Singh, G. S. Virk and J. Salo, 'Sharing of fake news on social media: Application of the honeycomb framework and the third-person effect hypothesis,' *Journal of Retailing and Consumer Services*, vol. 57, p. 102 197, 2020, ISSN: 0969-6989. DOI: `https://doi.org/10.1016/j.jretconser.2020.102197`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S0969698920306433`.

[45] S. Talwar, A. Dhir, P. Kaur, N. Zafar and M. Alrasheedy, 'Why do people share fake news? associations between the dark side of social media use and fake news sharing behavior,' *Journal of Retailing and Consumer Services*, vol. 51, Nov. 2019.

[46] K. M. Helkala and C. F. Rønnfeldt, 'Understanding and gaining human resilience against negative effects of digitalization,' in *Cyber Security: Critical Infrastructure Protection*. Cham: Springer International Publishing, 2022, pp. 79–91, ISBN: 978-3-030-91293-2. DOI: `10.1007/978-3-030-91293-2_4`. [Online]. Available: `https://doi.org/10.1007/978-3-030-91293-2_4`.

[47] Pew Research Center, 'Social media and news fact sheet,' en-US, Washington, D.C., Tech. Rep., Sep. 2022. [Online]. Available: `https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/`.

[48] E. G. Sivertsen, N. Hellum, B. A. and L. B. Bjørnstad, 'Hvordan gjøre samfunnet mer robust mot uønsket påvirkning i sosiale medier,' 2021. [Online]. Available: `https://www.ffi.no/publikasjoner/arkiv/hvordan-gjore-samfunnet-mer-robust-mot-uonsket-pavirkning-i-sosiale-medier`.

[49] Z. Sanderson, M. A. Brown, R. Bonneau, J. Nagler and T. J. T., *Twitter flagged donald trump's tweets with election misinformation: They continued to spread both on and off the platform.* 2021. DOI: `https://doi.org/10.37016/mr-2020-77`. [Online]. Available: `https://misinforeview.hks.harvard.edu/article/twitter-flagged-donald-trumps-tweets-with-election-misinformation-they-continued-to-spread-both-on-and-off-the-platform/`.

[50] T. K. Koch, L. Frischlich and E. Lermer, 'Effects of fact-checking warning labels and social endorsement cues on climate change fake news credibility and engagement on social media,' *Journal of Applied Social Psychology*, vol. n/a, no. n/a, DOI: `https://doi.org/10.1111/jasp.12959`. eprint: `https://onlinelibrary.wiley.com/doi/pdf/10.1111/jasp.12959`. [Online]. Available: `https://onlinelibrary.wiley.com/doi/abs/10.1111/jasp.12959`.

[51] F. Sharevski, R. Alsaadi, P. Jachim and E. Pieroni, *Misinformation warning labels: Twitter's soft moderation effects on covid-19 vaccine belief echoes*, 2021. arXiv: `2104.00779 [cs.SI]`.

[52] A. A. A. Ahmed, A. Aljabouh, P. K. Donepudi and M. S. Choi, *Detecting fake news using machine learning : A systematic literature review*, 2021. arXiv: 2102.04458 [cs.CY].

[53] B. Riedel, I. Augenstein, G. P. Spithourakis and S. Riedel, *A simple but tough-to-beat baseline for the fake news challenge stance detection task*, 2018. arXiv: 1707.03264 [cs.CL].

[54] F. Gereme, W. Zhu, T. Ayall and D. Alemu, 'Combating fake news in "low-resource" languages: Amharic fake news detection accompanied by resource crafting,' eng, *Information (Basel)*, vol. 12, no. 1, p. 20, 2021, ISSN: 2078-2489.

[55] S. Kreps, R. M. McCain and M. Brundage, 'All the news that's fit to fabricate: Ai-generated text as a tool of media misinformation,' *Journal of Experimental Political Science*, vol. 9, no. 1, pp. 104–117, 2022. DOI: 10.1017/XPS.2020.37.

[56] J. A. Goldstein, S. Chao Jason Grossman, A. Stamos and M. Tomz, 'Can ai write persuasive propaganda?,' 2023. [Online]. Available: https://osf.io/preprints/socarxiv/fp87b/.

[57] S. Dathathri, A. Madotto, J. Lan, J. Hung, E. Frank, P. Molino, J. Yosinski and R. Liu, *Plug and play language models: A simple approach to controlled text generation*, 2020. arXiv: 1912.02164 [cs.CL].

[58] 'Guide to social media training with somulator,' Accessed: 2023-05-25.

[59] J. Cito, V. Ferme and H. Gall, 'Using docker containers to improve reproducibility in software and web engineering research,' Jun. 2016, pp. 609–612, ISBN: 978-3-319-38790-1. DOI: 10.1007/978-3-319-38791-8_58.

[60] J. Kiesel, M. Mestre, R. Shukla, E. Vincent, D. Corney, P. Adineh, B. Stein and M. Potthast, 'Data for pan at semeval 2019 task 4: Hyperpartisan news detection,' 2019.

[61] LanguageTool-org, *Languagetool*, https://github.com/languagetool-org/languagetool, 2016.

[62] NbAILab, *Norwegian parliament speeches*, 2021.

[63] UiO. [Online]. Available: https://nettskjema.no/.

[64] S. Sütterlin, R. Lugo, T. Ask, K. Veng, J. Eck, J. Fritschi, M. Talha-Özmen, B. Bärreiter and B. Knox, 'The role of it background for metacognitive accuracy, confidence and overestimation of deep fake recognition skills,' *Lecture Notes in Computer Science*, vol. 13310, pp. 103–119, Feb. 2022.

[65] T. Kuzman, *Comparison of genre datasets: CORE, GINCO and FTD*, https://github.com/TajaKuzman/Genre-Datasets-Comparison, 2022.

[66] L. M. Tenstad and O. H. H., 'Maritim militærtrening i en usikker verden: En studie av sammenhengen mellom personlig hardførhet, opplevd mestringstro, kognitiv fleksibilitet og beslutningsstil i en maritim kontekst,' no, M.S. thesis, University of Bergen, 2022. [Online]. Available: https://bora.uib.no/bora-xmlui/bitstream/handle/11250/2999927/Masteroppgave-19-05-22.pdf?sequence=1&isAllowed=y.

[67] C. Leiter, R. Zhang, Y. Chen, J. Belouadi, D. Larionov, V. Fresen and S. Eger, *Chatgpt: A meta-analysis after 2.5 months*, 2023. arXiv: 2302.13795 [cs.CL].

[68] M. Tars, T. Purason and A. Tättar, 'Teaching unseen low-resource languages to large translation models,' in *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 375–380. [Online]. Available: https://aclanthology.org/2022.wmt-1.33.

[69] P. E. Kummervold, J. De la Rosa, F. Wetjen and S. A. Brygfjeld, 'Operationalizing a national digital library: The case for a Norwegian transformer model,' in *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, Reykjavik, Iceland (Online): Link"oping University Electronic Press, Sweden, 2021, pp. 20–29. [Online]. Available: https://aclanthology.org/2021.nodalida-main.3.

[70] Forsvarets Forskningsinstitutt, *Slik skal ffi forske på påvirkningsoperasjoner*, https://www.ffi.no/aktuelt/nyheter/slik-skal-ffi-forske-pa-pavirkningsoperasjoner, Accessed: 2022-04-22.

Appendix A

First Iteration

Here the survey-results from the first iteration of the within-subject design is presented. This work was done to evaluate and develop the models and pipelines.

Nettskjema

Spørreundersøkelse om datagenerert tekst

Oppdatert: 30. mai 2023 kl. 23:01

Denne spørreundersøkelsen er del av et prosjekt hvor vi ønsker å se nærmere på hvordan kunstig intelligens kan brukes for å generere tekst som oppfattes som autentisk i sosiale medier.
Nedenfor vil du bli presentert for flere korte tekster, din jobb vil være å ta stilling til om teksten er skrevet av et menneske eller en maskin.

Menneske eller maskin?

Antall svar: **31**

Svar	Antall	% av svar	
Vet ikke	2	6.5%	6.5%
Maskin	21	67.7%	67.7%
Menneske	8	25.8%	25.8%

Menneske eller maskin?

Antall svar: **31**

Svar	Antall	% av svar	
Vet ikke	3	9.7%	9.7%
Maskin	15	48.4%	48.4%
Menneske	13	41.9%	41.9%

Menneske eller maskin?

Antall svar: **31**

Svar	Antall	% av svar	
Vet ikke	3	9.7%	9.7%
Maskin	5	16.1%	16.1%
Menneske	23	74.2%	74.2%

Menneske eller maskin?

Antall svar: **31**

Svar	Antall	% av svar	
Vet ikke	7	22.6%	22.6%
Maskin	12	38.7%	38.7%
Menneske	12	38.7%	38.7%

Menneske eller maskin?

Antall svar: **31**

Svar	Antall	% av svar	
Vet ikke	5	16.1%	16.1%
Maskin	10	32.3%	32.3%
Menneske	16	51.6%	51.6%

Menneske eller maskin?

Antall svar: **31**

Svar	Antall	% av svar	
Vet ikke	1	3.2%	3.2%
Maskin	17	54.8%	54.8%
Menneske	13	41.9%	41.9%

Menneske eller maskin?

Antall svar: **31**

Svar	Antall	% av svar	
Vet ikke	3	9.7%	9.7%
Maskin	16	51.6%	51.6%
Menneske	12	38.7%	38.7%

Menneske eller maskin?

Antall svar: **31**

Svar	Antall	% av svar	
Vet ikke	3	9.7%	9.7%
Maskin	20	64.5%	64.5%
Menneske	8	25.8%	25.8%

Menneske eller maskin?

Antall svar: **31**

Svar	Antall	% av svar	
Vet ikke	1	3.2%	3.2%
Maskin	21	67.7%	67.7%
Menneske	9	29%	29%

Hvor sikker har du vært på dine svar?

Antall svar: **31**

Svar	Antall	% av svar	
5 - sikker	0	0%	0%
4	5	16.1%	16.1%
3 - hverken eller	12	38.7%	38.7%
2	8	25.8%	25.8%
1 - usikker	6	19.4%	19.4%

Var det noen tegn du så etter i tekstene for å ta ditt valg?

- Feil i ordbruk, faktafeil (Brukte ordet Statoil, som er gammelt)
- Flere steder var det innslag av dansk bed bruk av ordet «der» på en måte som vi ikke gjør på norsk.
- "unaturlig" grammatikk og bruk av ord
- Dårlig sammenheng på skrivefeil og tekst, menneske. Rar formatering eller utbytting av enkeltord, maskin. Bra norsk, vanskelig å si. Fargerikt språk, antakelig menneske.
- Tekst som virker oversatt fra et annet språk
- Nei
- Skrivefeil. Likevel er det vanskelig å si om skrivefeil er grunnet maskin som direkte oversetter om det er menneskelige feil.
- Syntaks
- Konsistent / ukonsistent språk
- Ordstilling
- Unaturlig setningsoppbygging, ord og begreper i malplassert kontekst, generell "dårlig" norsk, tekst som bærer preg av "engelsk gramatikk" (som det kan antas de mest sofistikerte språkgeneratorene er baserte på). Det er vanskelig å si, og det kan være vanskelig å skille mellom en kronglete formidling og maskingenerert tekst, men jeg synes i utgangspunktet at alle disse tekstene virket preget av noe "kunstig". Det er et intuitivt element også; mennesker og forfattere som bryr som om kunsten sin finner fargerike måter å uttrykke seg på, fordi de behersker nyansene i språket. Siden enhver dyktig forfatter har et visst særpreg på måten de uttrykker seg på, er det nok fremdeles utfordrende å trene ML algoritmer på å etterligne "kunst" feilfritt. Men igjen, jeg kan jo ha tatt feil på alle disse eksemplene... trodde alle var maskingenererte.
- Ordvalg & tegnsetting
- Skrivefeil som var så ut som noen som har gått «livets hare skole» kunne ha gjort.
- Ord som teknisk sett er riktige, men som virker kunstige (sfære f.eks)
- Tegnsetting, bruk av små og store bokstaver, og logisk oppbygging.
- Flyt, ordvalg, oppbygning av setninger
- Om det ga mening. Jeg tror ikke folk forventer at Bjørvika blir veldig anerledes hvis Høyre mistet makten (og ingen sier PÅ Bjørvika). Det er rart å debutere som artist i MGP-finalen, da må det ha skjedd i en semifinale. Og noen av skrivefeilene var så langt fra hvordan skrivefeil pleier å være
- Skrivefeil og setningsoppbygning.
- Ikke småfeil, men rare grammatikalske konstruksjoner kan kanskje indikere maskingenerert tekst.
- Statisk språk og utdaterte måter å skrive på.
- Veldig usikker om de tekstene som har blitt skrevet riktig. De kan være både AI eller menneske. Men de tekstene som er dialekt eller dårlig språk, tenker jeg at er skrevet et menneske.
- Skrivefeil, unaturlige setningser
- Store og små bokstaver, skrivefeil,

Appendix B

Second Iteration

Here is the survey distributed for the second iteration of the within-subject design. Here the models used were the fine-tuned models.

Nettskjema

Spørreundersøkelse om datagenerert tekst 2

Oppdatert: 30. mai 2023 kl. 23:03

Denne spørreundersøkelsen er laget for å se nærmere på hvor godt en maskinlæringsmodell kan generere tekst som kan fremstå som naturlig i sosiale medier. Nedenfor vil du presenteres for ulike tekster, din oppgave er å svare om du tror teksten er skrevet av et menneske eller en maskin.

Tekstene under kan fremstå som støtende for noen. Vi ønsker å presisere at tekstene under ikke er representative for deltakerne i prosjektets arbeid, men at ønsket har vært å se på polariserende tema innen sosiale medier, og hvordan maskinlæringsmodeller kan prestere innen dette området.

Menneske eller maskin?

Antall svar: **31**

Svar	Antall	% av svar	
Usikker	1	3.2%	3.2%
Menneske	7	22.6%	22.6%
Maskin	23	74.2%	74.2%

Menneske eller maskin?

Antall svar: **31**

Svar	Antall	% av svar	
Usikker	1	3.2%	3.2%
Menneske	17	54.8%	54.8%
Maskin	13	41.9%	41.9%

Menneske eller maskin?

Antall svar: **31**

Svar	Antall	% av svar	
Usikker	1	3.2%	3.2%
Menneske	18	58.1%	58.1%
Maskin	12	38.7%	38.7%

Menneske eller maskin?

Antall svar: **31**

Svar	Antall	% av svar	
Usikker	0	0%	0%
Menneske	9	29%	29%
Maskin	22	71%	71%

Menneske eller maskin?

Antall svar: **31**

Svar	Antall	% av svar	
Usikker	3	9.7%	9.7%
Menneske	12	38.7%	38.7%
Maskin	16	51.6%	51.6%

Menneske eller maskin?

Antall svar: **31**

Svar	Antall	% av svar	
Usikker	2	6.5%	6.5%
Menneske	24	77.4%	77.4%
Maskin	5	16.1%	16.1%

Menneske eller maskin?

Antall svar: **31**

Svar	Antall	% av svar	
Usikker	2	6.5%	6.5%
Menneske	10	32.3%	32.3%
Maskin	19	61.3%	61.3%

Menneske eller maskin?

Antall svar: **31**

Svar	Antall	% av svar	
Usikker	0	0%	0%
Menneske	20	64.5%	64.5%
Maskin	11	35.5%	35.5%

Hvor sikker var du på dine svar?

Antall svar: **5**

Svar	Antall	% av svar	
5 - Helt sikker	0	0%	0%
4	1	20%	20%
3 - Hverken eller	3	60%	60%
2	1	20%	20%
1 - Helt usikker	0	0%	0%

I prosessen med å velge, var det noen konkrete trekk som styrte valget ditt?

- Valg av ord
- Skrivefeil, store/små bokstaver, setningsoppbygning, noe innhold
- Verbbøying, store/små bokstaver, muntlig/uformell skriftlig tekst, kontekst i setningen
- Setninger der noe ikke helt gir mening, eller språket ikke virker helhetlig, gjorde at jeg valgte maskin. Virker det for mye som et troll som ikke kan skrive blir det menneske
- Liten forbokstav, rare oppbygninger og kommafeil
- Dårlig språk, grammatikk (feil eller konsistens)
- Unaturlige a-endelser flere steder.
- Språk, tegnsetting og setnibgsoppbygging. Min forståelse er at det er fornuftige folk som skal ha skrevet de menneskelige tekstene, ikke det generelle kommentarfelt på Internett. Dersom det er feil inngangsverdi, kan alle svarene byttes om. Antar maskinteksten er generert fra data fra kommentarfelt/internett
- Språket. Der det var litt avansert tenkte jeg det var maskin.
- Dette var vanskelig. Jeg tenkte at de påstander som hadde skrivefeil var skrevet av mennesker. De som så riktig ut kunne være begge deler og dermed valgte jeg "usikker".
- Donald Biden. Setningsoppbygning og detaljer i kontekst
- Jeg prøvde å "lytte" på teksten for å høre om dette er noe som hørtes "menneskelig" ut. Men, jo nøyere jeg leste dem, jo vanskeligere ble det å avgjøre. Valgte derfor å gå på førsteinntrykk og intuisjon.
- Rene ord feil (Donald Biden), rar norsk ordstilling og feil logikk (Afganistah/Balkan)
- Ganske sikker
- setningsoppbygging
- Jeg mener at maskinen lager tyngre setninger
- Setningsoppbygging, A endelsen, tegnsetting
- Fokusert på skrivefeil som jeg antyder vil komme fra et menneske. Enkelte benyttede begreper (f.eks gender) fikk meg til å tro at det er maskin
- Stor bokstav i starten av en setning, ord som ikke er like vanlig å bruke/bærer preg av å kunne være direkte oversatt fra et annet språk, skrivefeil
- A-endinger, rask avslutning på setningen

Appendix C

2x2 Factorial design survey

Here is the survey-results from the 2x2 Factorial design testing, where texts were tested in a general and domain-specific context.

Nettskjema

Evaluering av datagenerert tekst på ulike språk

Oppdatert: 31. mai 2023 kl. 22:34

Denne spørreundersøkelsen er laget for å evaluere i hvilken grad ulike språkmodeller presterer på ulike språk, og med ulike fokusområder. Målet er å undersøke hvordan disse språkmodellene kan fremstå som ekte brukere i sosiale medier.

Nedenfor vil du presenteres for en rekke tråder med ulikt fokus. Noe vil være mer generisk som en brukers feed, mens andre deler vil være mer målrettet (e.g. tekster i forbindelse med en hashtag). Din oppgave vil være å identifisere tekstene du mener er skrevet av en maskin.

Kjønn?

Antall svar: **24**

Svar	Antall	% av svar	
Ønsker ikke å oppgi	0	0%	0%
Annet	0	0%	0%
Mann	12	50%	50%
Kvinne	12	50%	50%

Din alder?

Antall svar: **24**

Svar	Antall	% av svar	
Ønsker ikke å oppgi	0	0%	0%
Over 65	2	8.3%	8.3%
56-65	3	12.5%	12.5%
46-55	3	12.5%	12.5%
36-45	1	4.2%	4.2%
26-35	14	58.3%	58.3%
16-25	1	4.2%	4.2%
Under 16	0	0%	0%

Hvilke av tekstene er skrevet av en maskin?

Antall svar: **23**

Svar	Antall	% av svar	
Polen og Slovakia gjør et nytt forsøk på å lure vestlige land som Norge til å sende jagerfly inn i Ukraina. Om regjeringen selger/sender våre F-16 inn i Ukraina så vil disse kunne brukes langt inn i Russland – vi vil stå ansvarlige for dette. Få heller i gang fredsforhandlinger!	4	17.4%	17.4%
Mens USA-NATO taper krigen de startet mot Russland, lager Kina fred mellom Iran og Saudi Arabia. Kina viser seg som den rake motsetning av vestens krigs-imperialisme. Fred og samarbeid heller enn konflikt og krig.	14	60.9%	60.9%
Eksperter, herunder representanter for IMF som jeg selv har snakket med, sier at det som nå har skjedd i Ukraina, kanskje var den alleste meste alvorlige politiske ulykke for Vladimir Putin i dette århundret.	11	47.8%	47.8%
Den bryter med alle prinsipper i FN-charteret og med folkeretten. Det er Russlands rett til å beskytte sine allierte i kampen mot et voldelige opprør som gjøres til en kriminell handling. Derimot kan vi ikke akseptere at den russiske militærmakten bruker alle midler for økt militarisering, i strid med Ukrainas territoriale integritet, også luftrom.	12	52.2%	52.2%

Hvilke av tekstene er skrevet av en maskin?

Antall svar: **24**

Svar	Antall	% av svar	
Du er født i en kropp og ikke med følelser. Følelser er en av de største årsakene til at mennesker ikke fungerer i samfunnet. Det at mennesker vrenger følelsene sine over til det umulige er skadelig for absolutt alle.	12	50%	50%
Er det noe som er absolutt sant? Og er det mulig å være absolutt lykkelig? Det var to spørsmål som jeg fikk som liten guttunge, og som har plaget meg siden. Nå har jeg kanskje funnet svaret på mitt første spørsmål.	8	33.3%	33.3%
Å tillate abort kun ved fare for mors liv betyr at du må bli fryktelig sjuk før du får helsehjelp. Gravide som står i fare for å dø fordi legen ikke tør å risikere fengelsstraff, slik ser endeholdeplassen for angrep på selvbestemmelse ut.	7	29.2%	29.2%
Jeg skal be deg om å løfte blikket nå. Jeg er selvfølgelig ikke motstander av at folk skaffer seg erfaring fra andre deler i verden eller at de studerer og reiser rundt. Men jeg mener at vi må ha en debatt om det er verdt det.	8	33.3%	33.3%

Hvilke av tekstene er skrevet av en maskin?

Antall svar: **23**

Svar	Antall	% av svar	
The people running our country are utter fools. These people are getting us on the brink of war with Russia. And I don't have any faith in them and I think most people don't	4	17.4%	17.4%
The way this country is run is so superior to any system that is anywhere else in the world because the checks and balances that were put in place by the Founding Fathers.	14	60.9%	60.9%
The United States of America is a nation of immigrants. We are a melting pot of people from all over the world who have come to this country to make a better life for themselves and their families. However, there are some people who do not agree with this. They believe that illegal immigrants should not be allowed to live in the U.S.	6	26.1%	26.1%
Equality of opportunity is the best way to achieve equality. If everyone is given the same opportunity to succeed, then no one will be held back by their race, gender, sexual orientation, religion, or anything else. However, if you give everyone an equal opportunity, some people will succeed and others will fail.	9	39.1%	39.1%

Hvilke av tekstene er skrevet av en maskin?

Antall svar: **23**

Svar	Antall	% av svar	
Several celebrities and Democrats, including Democrats serving on the J6 committee rose and spoke out to object certifying the 2016 election. That's not whataboutism. It's you're fine if we do it and don't know what a slippery slope is.	10	43.5%	43.5%
The people calling others “conspiracy theorists” still believe that Trump colluded with Russia to steal the 2016 election, that the Hunter Biden laptop story was “Russian disinformation”, that January 6th was an “insurrection”, and that the “vaccine” is safe and effective.	4	17.4%	17.4%
Podestas emails were also hacked and leaked, showing that the Clinton campaign had colluded with mainstream media outlets, such as CNN and The Washington Post, to push a false narrative about Trump s alleged ties to Russia and to smear Sanders as an anti-Semite.	7	30.4%	30.4%
Russia has been accused of meddling in the U.S. presidential election in an effort to help Donald Trump win the White House. The Kremlin has denied the allegations, and Trump has dismissed them as “fake news.” But Russia’s actions are having a real impact on the United States and the rest of the free world.	6	26.1%	26.1%

Hvor sikker er du på dine svar?

Antall svar: **24**

Svar	Antall	% av svar	
5 - sikker	0	0%	0%
4	4	16.7%	16.7%
3 - hverken eller	3	12.5%	12.5%
2	10	41.7%	41.7%
1 - usikker	7	29.2%	29.2%

Noen betraktninger du har gjort fra undersøkelsen du ønsker å dele?

- Ser for meg at enkelte grammatiske sammensetninger og/eller måter å uttrykke seg på, spesielt i forhold til dialekt og setningsoppnbygning slikt sett, kan være vanskelig for KI.
- Der mine egne synspunkter er enige med teksten er det enklere å godta dem og tro at de er skrevet av et menneske, og vice versa.
- Tror ikke maskiner skriver i jeg-form
- Tanken som slår meg er at jeg ikke har prøvd maskinskrevne tekster selv og derfor ikke helt vet hva jeg bør være obs på.
- Vanskelig å skille på hva som er datagenerert og ikke. Jeg tok meg selv i å trekkes mot at tekstene som virker å komme fra noen med far-right tankegang var datagenrerert- men det finnes jo slike folk også.
- Ilupilu
- Kontekst mangler - hva er tekstene et svar til?

Appendix D

Somulator tweets

Here are the tweets used in the Somulator experiment. Along with a description of the categories of each tweet, and if it is machine-generated or human-written.

rekkefølge:	Tekst:	sentiment:	menneske eller maskin:
1	Russerne er nødt til å hente sine styrker fra fengsler og leiesoldater uten moral. I Ukraina sliter de å ta imot alle som ønsker å bidra frivillig fra hele verden. Tydelig hvem som er de gode.	pro-ukrainsk	menneske
2	Zelenskyj bruker bare krigen for å fylle egne lommer. Han skryter på seg en seksdobling av økonomien, men det er bare juks, for det var jo ikke korrupsjon der før kuppet.	anti-ukrainsk	maskin
3	Hvorfor skal vi la Amerikanerne styre hvordan vi lever livene våre. Deres kapitalistiske tankesett har ikke gjort annet enn å skape problemer her i verden. Likevel oppfører seg som vår frelser.	anti-USA	menneske
4	Uten NATO ville vi aldri klart å støtte Ukraina i deres kamp mot tyranniet fra øst. Det er kun en grunn til at de svina ikke angriper hele Europa, og det er vår allianse.	pro-NATO	menneske
5	Vi burde gjøre mer som USA, som står på sitt og forsvarer sin egen suverenitet. Her lar vi alle andre hersje med oss, men USA de lar seg ikke pille på nesen. De skyter ned fiendens utstyr uten anger.	pro-USA	menneske
6	Å se ukrainske soldater ta ut russiske styrker vil for alltid være min yndlingsaktivitet. Deres overlegenet både i utstyr og taktikk viser tydelig at de gode vinner til slutt.	pro-ukrainsk	menneske
7	Før denne krigen var det korrupsjon og grådighet ute og gikk i Ukraina. Zelenskyj sa han skulle fikse opp i det, men han løy. Denne krigen er kun et spill for galleriet for å skjule all korrupsjonen.	anti-ukrainsk	menneske
8	FNs kvinnekommisjon vedtok å legge sitt neste møte til Russlands hovedstad, Moskva. Dette var en kraftig provokasjon overfor andre land. Denne skammelige avgjørelsen må aldri glemmes.	anti-russisk	maskin

9	Det er en skam at IOC og andre idrettsorganisasjoner skal stenge ut russiske idrettsstjerner fra å tjene penger til livets opphold. Bare fordi at Putin prøver å rydde opp i nabolandets problemer.	pro-russisk	menneske
10	Det skulle ikke være noe problem å sette inn noen tusen jagerfly dit for øyeblikkelig Å fjerne disse Russiske bombeflyene. Men gjør man det? Nei.	pro-ukrainsk	maskin
11	Det er bare positivt at vi sender forsvaret til å bidra i andre land som Afghanistan og Litauen. Det er viktig å forsvare frihet før den blir frarøvet her. Og våre soldater har godt av erfaringen.	pro-Forsvaret	menneske
12	I over 50 år har alliansen gitt oss forutsigbarhet og sikkerhet. NATO er en avgjørende del av vår trygghetspolitikk, omkranset av et transatlantisk bånd som prydes av gode politiske løsninger.	pro-NATO	maskin
13	Hvor mange ganger har ikke NATO vært den eneste grunnen til at vi har klart å deeskalere situasjoner, både i Asia og Europa. Uten NATO ville tredje verdenskrig vært kjempet nå.	pro-NATO	menneske
14	Vi vet at norske soldater deltar mer i internasjonale operasjoner enn noe annet land i verden, og det er ingen andre land som har så mange soldatliv på samvittigheten som Norge.	anti-Forsvaret	maskin
15	Nå har jeg fått nok av disse idiotene i forsvaret og våpenindustrien, som melker Norge for hver eneste krone som kunne gått til en bedre velferdsstat. De rike vil ha mer, og fanden vil ha fler.	anti-Forsvaret	menneske
16	En forutsetning for fred er at vi stopper folkemordet pá det ukrainske folket. Vi kan ikke lenger sitte med hendene i fanget. De etniske russerne i Øst-Ukraine må få fred.	anti-ukrainsk	maskin
17	Irak-krigen har skapt et monster. I USA sier man at USA-alliansen kommer alltid først. Nå er det viktig at vi sikrer at Nato-medlemskapet ikke medfører løftebrudd.	anti-USA	maskin

18	Det er Ukrainas president som har det øverste ansvaret. Det er han som skal sørge for å holde ro og orden i eget land. Slik fungerer et demokrati basert på folkestyre. Han viser manglende handlekraft.	anti-ukrainsk	maskin
19	Folk har hatt lett for å kritisere NATO når de har hjulpet folk i fjerne land i midtøsten. Men det er ikke like lett å kritisere dem når de må hjelpe til i vårt nabolag.	pro-NATO	menneske
20	I Norge ender vi opp med ministere vi absolutt ikke vil ha fordi demokratiet vårt har så mange skjulte avtaler mellom partiene. Vi burde se til USA som ikke driver med disse tåpelige hestehandlene.	pro-USA	menneske
21	Det er etter min mening nærmest uforståelig at Biden-administrasjons politikk på dette området ikke er blitt møtt av større internasjonal fordømmelse og kritikk.	anti-USA	maskin
22	I Russland har man ingen mulighet til å komme seg ut av fattigdom om man er født inn i det. Der er det kun oligarkene som får leve som konger, mens resten av landets innbyggere sulter.	anti-russisk	menneske
23	Så var vi nok en gang kommet til en tid hvor forsvaret maser om mer penger. De burde se til å skjerpe seg, og heller bruke pengene sine fornuftig. Kunne jeg valgt hadde de fått null.	anti-Forsvaret	menneske
24	Fordi vi gjennom vårt samarbeid med Iran, Irak og Syria i opprettelsen av Den islamske stat har destabilisert Syria. Det gir Russlands frykt om at Iran skal følge etter, kanskje mer legitimitet.	pro-russisk	maskin
25	Det er viktig med et forsvar som kan være en motmakt til stormakter som USA og Russland også i fredstid. Det er det våre soldater forbereder seg på.	pro-Forsvaret	maskin
26	De siste årene har vist hvorfor vi trenger NATO. Det ville ikke bare vært Ukraina som ble invadert om NATO ikke eksisterte. Det trengs et verdenspoliti, og NATO har de rette verdiene for den rollen.	pro-NATO	menneske

27	Det er flere ting som skiller den sittende administrasjonen fra Trump-administrasjonen. Denne administrasjonen står for en markant ny vektlegging av konflikten mellom israelere og palestinere	pro-USA	maskin
28	Uansett hvordan du vrir og vender på det, var det Russland som først brøt alle avtaler. Hvis Ukraina ikke hadde reagert med en gang, hadde det kommet flere russere.	anti-russisk	maskin
29	Det er bare økonomisk krigføring som gjelder. Derfor er USA den store stygge ulven, mens vi er de snille gutta. Jeg lurer på: Hvor er Bidens støtte til det folkemordet som pikene står midt oppe i der?	anti-USA	maskin
30	Russland har all rett til å ta tilbake Krim. – Men faktum er at Russland ikke har fått noe. Derfor er det reelt sett dette som nå skjer i Ukraina: en russisk invasjonsøvelse i Europa.	pro-russisk	maskin
31	Det er ingen som har gjort så mye for å forbedre vår kompetanse innen fysikk, matte og forståelse av verden som Russland. Vi burde ikke skyve dem vekk bare på grunn av interne problemer.	pro-russisk	menneske
32	NATO er ikke nødvendig for økonomisk vekst, ei heller for demokrati eller for fred. Og så skal man konsentrere seg om forsvar av Europa.	anti-NATO	maskin
33	Russere er kun superortodokse kristne som hater alt moderne. Det er ikke rart det går så dårlig i landet der, slik som de undertrykker kvinner og minoriteter. De gamle gubbene der vet ikke bedre.	anti-russisk	menneske
34	Ledere som Zelenskys politikk og innstilling til Russland også er et klart signal til Moskva om at NATO er en mer offensiv organisasjon en det man liker	anti-ukrainsk	maskin
35	Hvorfor får NATO lov til å holde på som de gjør, uten at det får konsekvenser. De er krigsprofitører som hauser opp ukrainerne kun for å tjene penger på deres lidelse.	anti-NATO	menneske

36	Se hvordan Ukraina samarbeider med blackrock og WEF for å privatisere Ukraina etter krigen. Denne krigen er ikke noe annet enn et forsøk fra ukrainske myndigheter på å tjene penger på de uskyldige.	anti-ukrainsk	menneske
37	Det er tydelig at folk har glemt hvordan verden så ut før NATO. Det er takket være dem at vi har fått en trygghet i hverdagen, hvor vi kan gjøre verden til et bedre sted, og samles som et folk.	pro-NATO	menneske
38	Det er klart av og til at man kanskje ikke har gjort den helt store rekrutteringsjobben. Jeg tror Forsvarssjefen også må ta sin del av skylden for det som er skjedd her.	anti-Forsvaret	maskin
39	Zelenskyj har ikke ett eneste positivt tiltak for Ukraina, men han øser penger inn i sitt eget luksusliv. Krigsgjelden, gasspriser, energi – alt blir dyrere.	anti-ukrainsk	maskin
40	Zelenskyj har kanskje en bakgrunn som skuespiller, men har vist at han fortjener å være president mer enn noen andre. Han kjemper for det han tror på.	pro-ukrainsk	menneske
41	Ukraina er et korrupt land, det er bare bra at Russland rydder opp der, slik at det kan bli et demokrati i Ukrania. Russerne omtaler seg som redningsmenn i Ukraina.	pro-russisk	maskin
42	Å se Vitalij Klitsjko forsvare Kyiv sammen med sine landsmenn er noe av det mest rørende jeg har sett. En mann som hadde midlene til å flykte, men som likevel har blitt værende for å ta vare på sine.	pro-ukrainsk	menneske
43	Hver eneste dag ser vi at Russland er villig til å inngå avtaler med lyssky personer, og henter ut terrorister for å støtte i sin krig. De har ingenting i verdenssamfunnet å gjøre.	anti-russisk	menneske
44	USAs folkevalgte er villig til å ta ekstra risiko med egen økonomi for å redde freden her i Europa. Vi står i evig gjeld til dem, som tar vår fred så seriøst.	pro-USA	menneske
45	Gaddafi fikk gjennomgå og ble slept gjennom gatene i sitt eget land fordi han ikke ville bøye seg for USA og godta deres dollar i eget land. Bøller er de!	anti-USA	menneske

46	Det er ikke lenge siden VG gjorde en reportasje om hvordan norske soldater liker å drepe mer enn de liker sex. Tror dere virkelig vi kan stole på at de skal forsvare oss?	anti-Forsvaret	menneske
47	Når vi ser hva Russland er villig til å gjøre mot Ukraina, Europas matkammer, da må vi begynne å forstå at vi ikke lenger kan vente. Vi må ruste opp Norge, før Russland får lyst på vår kystlinje.	pro-Forsvaret	menneske
48	Det er også en del land i NATO som ikke er så nøye på det, som Afghanistan f.eks. Der sier man at «nå skal vi bare bombe dem litt rundt omkring pga. terrorfaren".	anti-NATO	maskin
49	USA har løftet mange hundre millioner mennesker ut av fattigdom. De har en enorm evne til innovasjon, og de har bygd opp et tillitsforhold til både venner oghandelspartnere.	pro-USA	maskin
50	Vi må erkjenne at når man snakker om Russlands situasjon, eller nært sagt: nedlatende omtaler situasjonen i Russland fordi det fins noen få brudd, så tjener det ene formål å rettferdiggjøre vesten.	pro-russisk	maskin
51	Biden er et symptom på alt som er galt med USA. Hvordan kan man forsvare å bytte en kvinnelig basketspiller som røyket hasj mot en fengslet våpenhandler kjent for sine grusomheter?	anti-USA	menneske
52	Vi er villig til å kaste bort mer penger på en helsetjeneste som knapt fungerer og en velferdsstat for snyltere. Men det å investere i et godt forsvar, det gidder vi ikke. Aldri mer 9.april	pro-Forsvaret	menneske
53	Russlands spesialoperasjon i Ukraina er kun for å skape fred og ro i et område som tilhører Russland. Likevel beskrives det hele som en krig. Kun fordi verden er redd for at Russland skal lykkes.	pro-russisk	menneske
54	I motsetning til mange av Republikanerne her i salen føler jeg at det er en viss forbedring i den amerikanske administrasjonens håndtering av Nord-Korea.	pro-USA	maskin

55	For de som er skeptiske til at vi burde bruke så mye penger på Forsvaret så vil jeg kontre med at vi burde bruke mer. Det er de som har vært i Forsvaret som driver Norge fremover.	pro-Forsvaret	menneske
56	Jeg trodde det ikke kunne bli verre etter at Trump tok over det ovale kontor, men det kunne det visst likevel. De har sluttet å bruke gode ledere, og heller valgt pedofile og senile.	anti-USA	menneske
57	Disse amerikanerne burde læres opp i litt historie og geografi. De tror de er viktige fordi de ikke blir vist annet enn sitt eget speilbilde, men det finnes en hel verden utenfor gjerdene deres.	anti-USA	menneske
58	I Russland er det øverste sjiktet dominert av russere, mens det i vest er dominans av innvandrere. Da får du en annen dynamikk, fordi du faktisk har et fremmed element i samfunnet ditt.	pro-russisk	maskin
59	Nord stream ble ødelagt av russerne, enten med vilje eller av deres inkompetanse. Men istedenfor å forbedre seg skylder de på Norge og USA. De klarer aldri ta ansvar for egne feil.	anti-russisk	menneske
60	Det er litt underlig å høre den samme argumentasjonen hver gang man diskuterer med Putin-tilhengere eller -apologeter: Man må ikke kritisere Russland.	anti-russisk	maskin
61	Når Russland taper krigen og verden styres av den kapitalistiske elite kan dere takke dere selv. Folks forakt mot russere gjør meg kvalm. De står opp mot urett, og behandles som fienden.	pro-russisk	menneske
62	Gunnar Sønsteby og Max Manus måtte klare seg på egenhånd, mens vårt eget forsvar lå med brukket rygg på grunn av elendige politikere. Vi kan ikke gjøre den samme feilen igjen. Vi må gjøre ruste opp.	pro-Forsvaret	menneske
63	Når skal Norge innse at vi er på feil side av historien når vi er på lag med NATO. Gang på gang har våre soldater blitt brukt på å bombe uskyldige mennesker i fjerne land, så oljeprisene kan stige.	anti-NATO	menneske

64	Jeg ser ingen grunn til at Norge skal støtte forslag som bare har til hensikt å ramme ett spesielt land, og som rammer Ukraina omtrent i samme grad som en atombombe.	pro-ukrainsk	maskin
65	De ukrainske soldatene sa de skulle spille fotball med hodene til tsjetsjenerne. Kan dere ikke se hvor grusomme ukrainerne også er? De er ikke noe bedre enn russerne dere henger ut i media.	anti-ukrainsk	menneske
66	Det ble klart uttrykt i NATOs hovedkvarter at det er svært viktig at vi får et raskt og godt utfall av dette. Mange av soldatene blir nok drevet ut av byen av de pro-russiske demonstrantene.	pro-ukrainsk	maskin
67	Tolkien sa "ondskap kan aldri skape, bare gjøre korrupt". Det har Russland vist at stemmer. De skaper ingenting selv. De bare stjeler fra andre og ødelegger hva andre lager.	anti-russisk	menneske
68	Jeg er veldig bekymret for hvordan verden skal kunne holdes sammen, når vi ikke engang er i stand til som statsorganisasjon å holde styr på egen militærmakt.	anti-Forsvaret	maskin
69	For oss er det en selvfølge at Forsvarssjefen har siste ord når det gjelder hvordan vi bruker penger. Det betyr at pengene brukes på en måte som vil gagne Forsvaret	pro-Forsvaret	maskin
70	Zelenskyj er bare en skuespiller. Hver eneste dag bruker han mer tid på å snakke med Hollywood og USA, mens han lar sine landsmenn dø. Jeg håper Putin tar han snart.	anti-ukrainsk	menneske
71	Det er mulig det er meg det har klikka for, men vi kan ikke bare akseptere at vi én dag kan sitte igjen med Putin. Det er jævlig mange ting verdenssamfunnet kan gjøre overfor Russland.	anti-russisk	maskin
72	NATO-ledelsen gjør en av de dårligste jobbene du finner. Og det verste er de har ikke peiling på hvordan de skal løse de mest komplekse problemene, og har ingen respekt for menneskelige tap.	anti-NATO	maskin

73	Det er viktigere enn noen gang å fokusere på NATOs transatlantiske dimensjon. Dagens politiske utvikling i Europa er dramatisk og uten sidestykke i vår tid.	pro-NATO	maskin
74	Det er jo sånn at et land trenger et forsvarsbudsjett, og det budsjetteres jo i hver eneste regjering uavhengig av hvile partier som styrer. Det har ingenting med ideologi å gjøre.	pro-Forsvaret	maskin
75	Når russerne bomber Kyiv og andre storbyer i Ukraina så hjelper alle til for å ta vare på hverandre. Vi har glemt å gjøre det samme her i landet. Måtte Ukraina vinne, og vise oss veien. Slava Ukraini!	pro-ukrainsk	menneske
76	USA er og har alltid vært et fyrtårn for resten av verden. Uten USA ville verden rast for lenge siden. Bidens arbeid med å ta vare på sine borgere har også bært frukter utenfor USA.	pro-USA	menneske
77	I dag, spesielt med de store endringene som skjer i det norske samfunnet, ser vi at Forsvaret har en vesentlig plass også i framtidens kompetansesamfunn.	pro-Forsvaret	maskin
78	Men jeg tror ikke vi kan ha en situasjon der vi uthuler NATOs vedtak, eller der NATO-operasjoner mister sin troverdighet hvis det er enkelte medlemsland som ilegger seg selv særordninger.	anti-NATO	maskin
79	En felles utenrikspolitikk og utvikling av en fellesskapsfølelse, det være seg i EU, NATO eller i OSSE, er avgjørende. Bare slik kan Ukraina sikre stabilitet i sitt område.	pro-ukrainsk	maskin
80	Kvinnene i det norske Forsvar representerer halvparten av Norges befolkning. De bør være på lik linje med menn som vernepliktige, underlagt den samme lovgivning – og behandles deretter.	pro-Forsvaret	maskin
81	Hvem tror oppriktig på at NATO noen gang var lagd for å gjøre verden til et bedre sted. De ble skapt for å stoppe verden fra å bli et mer likestilt samfunn. For å stoppe sosial sikkerhet.	anti-NATO	menneske

82	Både soldater og befal, sivile som har hatt en tilknytning eller kontakt med soldatene, også barn, har blitt rammet. Er det en ting vi vet, er det at Forsvaret ikke ivaretar sitt etiske ansvar.	anti-Forsvaret	maskin
83	Det finnes mye bra å bruke pengene sine på, men forsvaret er ikke en av dem. De sløser bort penger i hytt og pine på båter de ødelegger i fredstid. Mye for sløseriombudsmannen å se på.	anti-Forsvaret	menneske
84	Når det gjelder frihet, er USA det desidert viktigste landet. Og nå snakker jeg ikke om økonomisk frihetsfølelse, men om muligheten for alle amerikanere til å leve i fri og uavhengig utfoldelse.	pro-USA	maskin
85	Synes ikke dere det er litt rart at Ukraina påstår de står i en kamp mot Goliat, når de får våpen og utstyr direkte fra den største Golia som finnes. Ukraina er den ekte synderen her.	anti-ukrainsk	menneske
86	NATO-styrkene som er utplassert i mange av de tidligere kommunistlandene, har én ting til felles: De har som oppdrag økt profitt for store amerikanske selskaper.	anti-USA	maskin
87	Presidenten har nå lagt fram en plan som er et uttrykk for stolthet, som viser at Ukraina står for noe ganske annet enn det som ligger til grunn for krisen i Ukrainia.	pro-ukrainsk	maskin
88	Jeg er glad for å se at Biden har fått USA på styr igjen. Et land som ikke lenger er fremmedfiendtlig, men som er villig til å ta imot mennesker med en drøm om å lykkes her i livet.	pro-USA	menneske
89	Det er altfor få kvinner i Forsvarets forskjellige ledd. Det er ikke slik at vi ønsker en svertekampanje mot menn, men det vi vil ha en debatt om er det som har med kultur og ledelse osv.	anti-Forsvaret	maskin
90	Vi så jo i forrige periode, da man hadde den kalde krigen, at det var USA som var den store, slemme ulven. Og Putin har til en viss grad snudd dette på hodet, og er en liten ulv i denne sammenheng.	anti-russisk	maskin

91	Alle som prøver å stoppe denne galskapen, møtes med ufattelig lite respekt også fra USA. Det verste er at mange av dem tror de forsvarer demokratiet, friheten, men hva er det egentlig de gjør?	anti-USA	maskin
92	Putin er en leder vi skulle hatt flere av. En leder som har vært "på gulvet" og jobbet seg oppover. Det er lenge siden vi så den slags ledere i andre land.	pro-russisk	menneske
92	Forsvaret vårt er bare bortkastet. Det har dødd flere i Forsvaret på norsk jord de siste tiårene enn det har dødd på operasjoner i utlandet. De er der bare til for å sløse penger.	anti-Forsvaret	menneske
94	Som mange har sagt, er det ikke USA, men europeerne som har vært den store stygge ulven. Jeg tror det er mye positivt amerikanerne har bidratt med. De har på mange måter vist vei for oss.	pro-USA	maskin
95	La oss slå det fast en gang for alle: NATO skaper faktisk færre overskrifter, konflikter, lidelser og død enn noen annen internasjonal organisasjon, men så gjør NATO også mer.	pro-NATO	maskin
96	NATO-toppene representerer en enorm kompetanse og erfaring. Den samlede erfaringsbakgrunnen er enorm, også innenfor den sikkerhetspolitiske sektor.	pro-NATO	maskin
97	For å bruke forsvarsminister Bjørn Arild Gram som eksempel: Han var på NATO-toppmøtet i Praha i fjor og sa at målet for NATO de neste 20 årene er ØST-EUROPA.	anti-NATO	maskin
98	NATO har to målsettinger. Økonomisk vekst i Nord-Amerika, og det andre er at USA skaper stabilitet i Europa. Derfor støtter NATO selvforsvarsprosjekter, som oppbygging av missilfri sone i Øst-Europa.	pro-NATO	maskin
99	Du kan si hva du vil, men hver eneste dag gjør det mer og mer tydelig at NATO ikke er til for å beskytte oss. De er de rikes private hær for å få det de vil, på bekostning av våre penger.	anti-NATO	menneske
100	Jens Stoltenberg og NATO bruker bare Ukraina for å slite ut Russland. Det er så tydelig at det eneste de ønsker er å fjerne sine motstandere for å kunne presse sin ideologi på resten av verden.	anti-NATO	menneske

NTNU

Kunnskap for en bedre verden