Karoline Bonnerud

Write Like Me

Personalized Natural Language Generation Using Transformers

Master's thesis in Computer Science
Supervisor: Björn Gambäck
June 2021

Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

NTNU
Norwegian University of
Science and Technology

Karoline Bonnerud

Write Like Me

Personalized Natural Language Generation Using Transformers

Master's thesis in Computer Science
Supervisor: Björn Gambäck
June 2021

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

NTNU
Norwegian University of
Science and Technology

Abstract

State-of-the-art language models with attention mechanisms, *transformers*, have revolutionized the field of natural language processing due to their demonstrated success within a variety of tasks. However, there are still numerous aspects to explore concerning the generation of natural language using transformers. At the same time, personalized open-ended natural language generation is attracting widespread interest. Hence, this thesis aims to combine personality psychology with state-of-the-art transformers to generate personalized open-ended short text for social media based on a fictive author's personality, age, and gender.

Two different transformers are compared on the task of personalized natural language generation, an autoregressive model and an autoencoding model, differing in their training procedures for learning language representations. Autoregressive models are trained by learning connections between which words often follow each other in sequences of text. On the other hand, autoencoder models learn language representation by repeatedly being exposed to texts where certain words are missing and then asked to figure out suitable words to fill the gaps.

This study is the first to compare several state-of-the-art transformers on the task of generating personalized natural language. It is also the first study applying the Big Five personality model to personalized natural language generation.

The results show that autoregressive language models are far more suitable for personalized natural language generation than autoencoding models. The autoregressive model obtains better results concerning fluency and coherence in generated texts and preserves characteristics of personality, age, and gender.

Notwithstanding, a lack of suitable automatic evaluation metrics is a significant drawback within the field of personalized natural language generation. No standard metrics are established, hindering comparable results and continuous development in the area. This study proposes and employs an automatic evaluation procedure based on the success of automatic personality prediction and author profiling.

This research is the first step towards enhanced personalized natural language generation, which is the foundation for obtaining extensive personal writing assistance in a wide range of domains.

Sammendrag

De nyeste og mest avanserte forhåndstrente språkmodellene med oppmerksomhetsbaserte dyp-læring-teknikker har revolusjonert feltet for språkteknologi. Slike språkmodeller har vist seg å være svært suksessfulle på en rekke oppgaver innen intelligent tekstanalyse og språkforståelse. Til tross for denne suksessen er det fortsatt mange aspekter tilknyttet disse modellene som må utforskes nærmere. Det også en økende interesse for personlig tilpasset språkgenerering. Derfor er formålet med denne studien å kombinere personlighetspsykologi med forhåndstrente språkmodeller for å generere korte tekster rettet mot sosiale medier, som er ment å etterligne skrivestilen til gitte personlighetstrekk, aldre og kjønn.

Denne studien sammenligner prestasjonen til to ulike avanserte språkmodeller når det gjelder å generere personlig tilpasset språk. Modellene er henholdsvis en *autoregressiv* modell og en *autoencoder* modell. Det som skiller dem fra hverandre er hvordan de er forhåndstrent for å lære seg representasjonen av språk. Autoregressive modeller er trent opp ved å lære sammenhenger mellom hvilke ord som ofte etterfølger hverandre. På den andre siden lærer de autoencodede modeller seg språkrepresentasjon ved å gjentatte ganger bli eksponert for tekster hvor enkelte ord er plukket ut og hvor modellen da blir bedt om å sette inn passende ord i hullene.

Dette er den første studien som sammenligner flere forhåndstrente språkmodeller med oppmerksomhetsbaserte dyp-læring-teknikker på generering av personlig tilpasset naturlig språk. Det er også den første studien innen personlig tilpasset språkgenerering som benytter femfaktormodellen for å representere personlighet.

Resultatene tilsier at autoregressive modeller er bedre enn autoencoder-modeller for personlig tilpasset språkgenerering. Den autoregressive modellen oppnådde bedre resultater både med hensyn til å generere grammatisk korrekt tekst og tekst som gir mening. Samtidig evner den autoregressive modellen også å generere tekster som bevarer karakteristikker for spesifikke personlighetstrekk, aldre og kjønn.

Til tross for dette er det en stor mangel på formålstjenlige metoder for å evaluere personlig tilpasset generert språk. Det medfører en betydelig ulempe innen feltet da det gjør det svært utfordrende å sammenligne resultater på tvers av studier ettersom man ikke er enige om hvilke metoder som bør benyttes for å måle prestasjoner. Denne studien foreslår og benytter en evalueringsprosedyre som er basert på suksesser innen automatisk prediksjon av personlighet og forfatteridentifisering.

Forskningen som er gjennomført er det første steget mot personlig tilpasset språkgenerering, som igjen er grunnlaget for intelligente, personlige tilpassede skriveassistenter. Denne studien er gjort på tekst fra sosiale medier, men personlig tilpasset språkgenerering kan overføres til alle domener.

Preface

This Master's Thesis is written as the final work of achieving the Master of Science in Computer Science degree from the Norwegian University of Science and Technology (NTNU) in Trondheim, Norway. The work has been supervised by Björn Gambäck and has been conducted within the Data and Artificial Intelligence Group at the Department of Computer Science.

I would like to express special thanks to my supervisor, Björn Gambäck, for his wonderful guidance and excellent feedback throughout the process of writing this thesis. I have enjoyed all our engaging conversations within language technology and computational linguistics. I have always been fascinated by language, and discovering the area of natural language processing and computational linguistics has been an inspiring journey.

I would also like to express gratitude to all my friends and family. Not to forget, my five years at NTNU would never have been the same without Abakus, the student union for Computer Science and Communication Technology. I am forever grateful for everyone I've gotten to know along the way.

Karoline Bonnerud
Trondheim, 11th June 2021

Contents

List of Figures

List of Tables

1 Introduction

The development of pretrained language models has revolutionized the field of natural language processing, including natural language generation. These models can produce text so fluent that it can be difficult to distinguish between text written by humans and text generated by the models. This thesis explores pretrained language models' ability to write coherent texts conditioned on a fictive author's personality, age, and gender for the social media domain.

This chapter will first describe the background and motivation behind the research. The goal and research questions are presented in Section 1.2, whereas the research method is described in Section 1.3. Important aspects to keep in mind regarding this thesis are covered in Section 1.4. Contributions are summarized in Section 1.5, lastly an overview of the upcoming chapters is given in Section 1.6.

1.1 Background and Motivation

Pretrained language models using attention mechanisms, *transformers*, were introduced by Vaswani et al. (2017) and have since been on the rise and gained great interest. In short, the attention mechanism enables the models to pay attention to relevant parts of the input when computing the output, hence focusing more on what is learned to be relevant. Pretrained language models with attention have performed remarkably well on natural language processing tasks and have revolutionized the field.

Although pretrained language models have shown significant improvement within natural language processing tasks like predicting the next word in a sentence and infilling missing words in a sentence, more intricate natural language generation tasks using pre-trained language models still lack research and remain not fully explored. One of these tasks is open-ended controllable text generation, which is still rising and receiving more awareness. Controllable generation denotes, for instance, controlling the writing style, the expressed emotions, and the thematical content of generated texts. This thesis aims to explore controllable personalized text generation within the social media domain with respect to personalization in terms of the writing style.

Social media platforms have connected humans across the globe. Two of the largest platforms, Facebook and Twitter, generate a massive amount of data every second. Twitter is a microblogging platform where users can post tweets consisting of text, including emoticons, hyperlinks, and mentions of other users. Facebook was primarily designed for users to connect with their friends and family. However, Facebook is today a complete platform for discovering news, advertisements, and other content not posted by family or friends.

The motivation behind this project is to investigate pretrained language models' ability to generate conditional personalized short texts for social media. Short texts in this context denote personal texts expected to be posted on Twitter and Facebook. The project will aim to achieve open-ended conditional text generation within the social media domain using high-level author attributes for controlling the writing style.

1.2 Goals and Research Questions

This Master's Thesis aims to unite the field of personality psychology and natural language generation by exploring personalized natural language generation.

Goal *Contribute to the field of personalized natural language generation by exploring methods for the generation of natural language for social media conditioned on a fictive author's personality.*

A fundamental question to be answered is with what level of certainty that gender, age, and personality traits can be inferred from text written on social media and what are linguistic characteristics for the different personalities, genders, and age groups. Hence the first research question.

Research Question 1 *How successful are state-of-the-art methods for automatic personality prediction of social media users?*

In exploring and deciding methods for the generation of personalized natural language for social media, differences between autoregressive and autoencoding language models will be examined. Methods will be considered suitable with respect to generating grammatical correct and coherent text and for preserving and incorporating personality, age, and gender in the generated texts.

Research Question 2 *What are suitable methods for generating personalized natural language?*

In the exploration of suitable methods, it is essential to evaluate and compare the generated texts. Hence appropriate methods for evaluating both the fluency and whether the personalization is successful must be in place.

Research Question 3 *What are suitable and efficient methods for evaluating personalized natural language generation systems?*

To summarize, the overall goal of the Master's Thesis is to explore methods for natural language generation of texts for social media that are conditioned on a fictive author's personal attributes, such as age, gender, and personality. The term social media text is meant to capture tweets and Facebook status updates posted by human users on the respective platforms Twitter and Facebook.

1.3 Research Method

Different research methods will be utilized to answer the three research questions. Towards gaining a sufficient understanding of the personalized natural language generation field and discovering potential gaps in existing research, there is a need to conduct a literature review on the topic.

An additional literature review on automatic personality prediction was also conducted as a part of a specialization project preparing for the Master's Thesis. Relevant findings with respect to this research will be synthesized and presented.

A complete system must be built for realizing experiments on personalized natural language generation, even though the system is not a goal in itself. The system is chosen to be built following a design and creation strategy, ensuring a systematic procedure that facilitates repeatability and quality. The implementation will be carried out by first building a working system prototype, then following a cycle of analysis, design, and implementation to reach a final system. When the system is developed, an experimental research method will be used. Experiments will be conducted according to an experimental plan, which will be created. Lastly, efficient and suitable evaluation methods will be used to evaluate the results.

1.4 Disclaimer

When researching personalized natural language generation, three important aspects are necessary to keep in mind. First and foremost, transformers are pretrained on a massive amount of unfiltered text and can, for that reason, produce text that can be perceived as offensive. When using such models in this project, there is no intention to harm, and the generated samples do not necessarily represent the meanings or intentions of the author.

Secondly, please note that two genders are used in this project because those are the gender categories represented in the existing datasets. Lastly, humans' personalities describe their tendencies to behave, think, and act in particular manners. Note that these are tendencies of behavior, not facts. Humans should not be placed and understood for a lifetime in fixed categories based on their measured personality traits.

1.5 Contributions

To summarize the thesis findings, the most outstanding contributions are the following:

- The design and implementation of a system using state-of-the-art language models for generating personalized natural language conditioned on personality, age, and gender.

- The finding that autoregressive language models are more suitable for natural text generation than autoencoding language models.

- A preparation and concatenation of the myPersonality dataset from 2013 and the PAN15 Author Profiling dataset, enabling them to be used together on natural language processing tasks.
- The identification of a need for an established baseline within personalized natural text generation to support development in the field and facilitating comparable results.

1.6 Thesis Structure

The rest of this Master's Thesis is organized in the following manner:

- Chapter 2 gives the necessary background theory to familiarize the reader with the relevant topics used in the thesis.
- Chapter 3 gives an introduction to the field of automatic personality prediction and author profiling.
- Chapter 4 covers a structured literature review and related work within the field of personalized natural language generation.
- Chapter 5 presents the myPersonality and PAN15 Author Profiling datasets which are to be used in the experiments.
- Chapter 6 describes the architecture designed and implemented to build a system for personalized natural language generation.
- Chapter 7 provides the experimental plan and the setup used in the experiments and presents the experimental results.
- Chapter 8 evaluates the obtained results and discusses the findings in light of their implications and the existing literature.
- Chapter 9 concludes the thesis in light of the research goal and questions and suggests further work within the field of personalized natural language generation.
- Appendix A consists of the structured literature review protocol for the literature review conducted on personalized natural language generation.
- Appendix B contains the quality assessment table of the structured literature review protocol in Appendix A.
- Appendix C has the structured literature review protocol of the literature review conducted on automatic personality prediction.
- Appendix D presents a subset of the generated personalized texts from the experiments.

- Appendix E shows the evaluation form used for the human assessment of generated texts.

2 Background Theory

This chapter will give the necessary background theory for understanding the research questions and provide context to the research conducted. First, Section 2.1 presents the Big Five personality model, which is used as the psychological basis for modeling personality in the experiments. Section 2.2 introduces deep learning and the development that has led to the transformer-based pretrained models used in this thesis.

Fundamentals of text processing are covered in Section 2.3. Continuing to Section 2.4, natural language processing and generation and related topics are explained. Note that Section 2.1 and Section 2.3 are revised and updated sections from the specialization project.

2.1 The Big Five Personality Model

The psychological field of personality is concerned with humans' personalities and how personality traits vary between individuals. The Big Five model is the most established for explaining human personality traits, and an introduction to the model is provided in this section.

The Big Five personality model is also known as the Factor-Five model or the OCEAN model. It describes human personality in five overall traits: Extraversion, Neuroticism, Agreeableness, Conscientiousness, and Openness to experiences. The model argues that these five dimensions can describe all human personality traits (McCrae and John, 1992; Goldberg, 1990). Each of the five traits is measured on a spectrum between two pairs of extremities. Figure 2.1 illustrates the extremities for each of the traits.

These pairs of extremities describing each factor are central aspects to make clear with the Big Five model. Extraversion describes whether people are quiet and reserved or outgoing and warm and is measured between *introverted* and *extroverted*. Hence introverted and extroverted denotes two opposite traits on the extraversion spectrum. Neuroticism looks at whether a person tends to behave calm and confident or more nervous and anxious and is measured between *stable* and *neuroticism*. The trait agreeableness is measured on the scale between *hostile* and *agreeable*, indicating the degree of kindness and trustfulness. Conscientiousness measures the dimension of preference for plans and preparations, giving a spectrum between *spontaneous* and *conscientious*. Lastly, openness describes the openness to experiences, ideas, and imagination and is measured on a scale from *closed* to *open*.

The Big Five personality traits can be quantified using different instruments. The Revised NEO Personality Inventory (NEO PI-R) is a personality test used to determine the five dimensions. In accordance, NEO PI-R includes six subcategories per personality

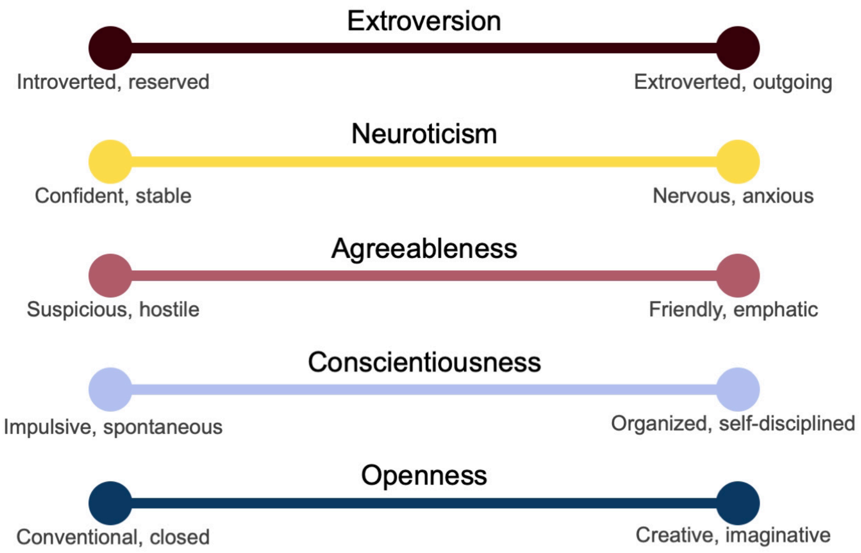


Figure 2.1: The traits of the Big Five personality model.

trait, giving even a more detailed description of the facets of the personality. The original inventory consists of 240 questions, whereas a shorter version with 60 questions also exists.

The Big Five Inventory (BFI-44) is another instrument that consists only of 44 short statements for self-reporting of personality traits. A version of the inventory with only ten questions to answer is also released, the Big Five Inventory 10 (BFI-10). Due to BFI-44 and BFI-10 having far fewer questions to answer than the NEO PI-R, those are considered more suitable when time is limited, and even BFI-10 is shown to achieve acceptable reliability and validity (Rammstedt and John, 2007).

2.2 Deep Learning

This section is provided to give an understanding of the advancements within deep learning related to the models used in the thesis, hence covering the path leading to transformer-based state-of-the-art language models. First, it is necessary to take a step back and start by examining the simplest type of neural network, feed-forward neural networks.

2.2.1 Feed-Forward Neural Networks

The aim of feed-forward neural networks, also denoted multilayer perceptrons, is to approximate a function. A feed-forward neural network is composed of perceptrons,

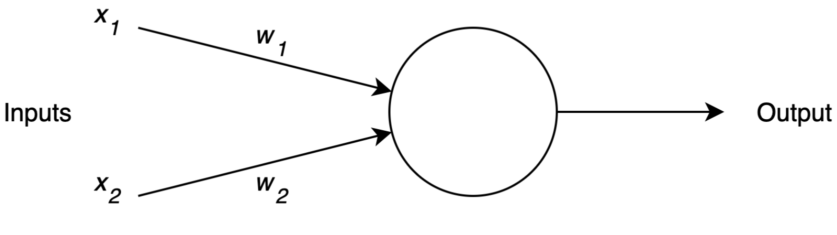


Figure 2.2: An illustration of a perceptron.

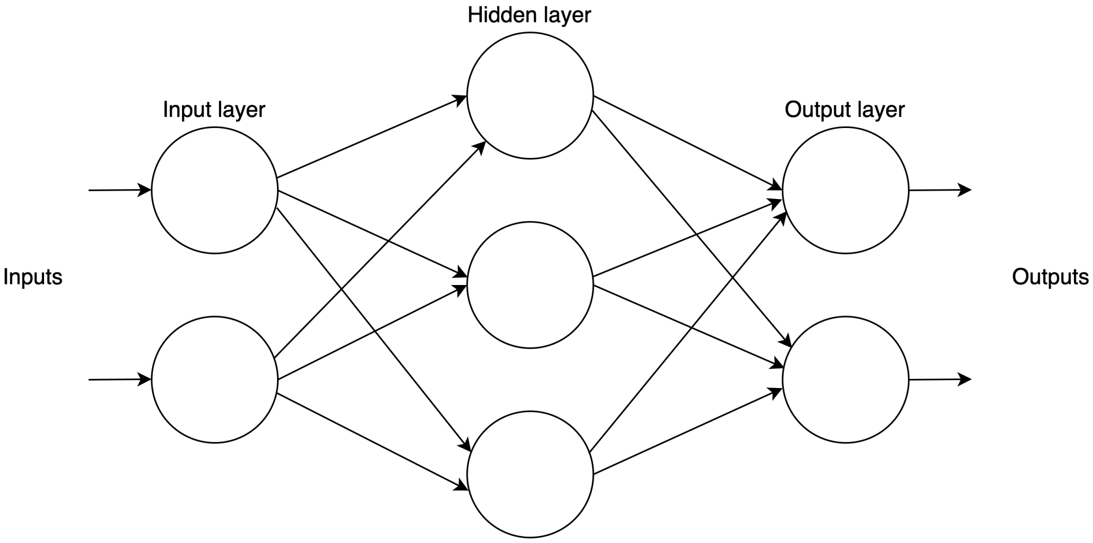


Figure 2.3: An illustration of a feed-forward neural network.

which are artificial neurons (Goodfellow et al., 2016). Figure 2.2 illustrates a perceptron in its simplest form. The perceptron takes in weighted inputs and uses a defined activation function to compute the output. The activation function describes how the perceptron is handling the input data and thus computes the output value. Multiple perceptrons organized in layers compose a feed-forward neural network. See this illustrated in Figure 2.32.3. The first layer of a feed-forward network is called *the input layer*, corresponding to the last layer is *the output layer*. All layers between are denoted as *hidden layers*.

As mentioned, the goal of a feed-forward neural network is to approximate a function with a minimum error by adjusting the weights in the network. For the network to know how to adjust the weights during training, a *loss function* is used to compute the difference between the current output and the desired output, and a *learning rate* sets how much the weights should be adjusted for each training step.

2.2.2 Recurrent Neural Networks

Recurrent neural networks (RNNs) (Goodfellow et al., 2016) extend feed-forward neural networks from Section 2.2.1 by adding loops that allow the network to use what it has learned in the past to compute the present. RNNs are especially suitable for handling sequential data such as text. Within natural language processing, sequences of words can capture textual semantics, and RNNs better preserve these because of their built-in memory. However, RNNs suffer from the vanishing gradient problem when processing long sequences. That is, over time, the gradient storing the sequential information will gradually be smaller and smaller, and hence information will disappear.

2.2.3 Long Short-Term Memory Networks

To overcome the vanishing gradient problem of RNNs from Section 2.2.2, but keeping the short-term memory, Long Short-Term Memory (LSTM) models were introduced by Hochreiter and Schmidhuber (1997). These networks consist of *cells* with three gates each. *The forget gate* is responsible for getting rid of the information the cells are going to forget, which is done by multiplying the actual positions by zero. New information to the cell is added via *the input gate*, and *the output gate* uses the information from the current cell state and output the value which should be passed to the next hidden state. These mechanisms make LSTM networks better at tasks requiring long-term dependencies to be remembered.

2.2.4 Sequence-to-Sequence Models

Sequence-to-Sequence (Seq-2-seq) (Sutskever, Ilya and Vinyals, Oriol and Le, Quoc V., 2014) models are applied for tasks where one sequence is transformed into another sequence. Examples of this are manifold in natural language processing, for instance language translations and text summarization. In both these tasks, a sequence of text is fed to a model, and the expected output is another meaningful sequence. A Seq-2-seq model is realized using an encoder and a decoder. According to its names, the encoder is responsible for encoding the input into a hidden vector representation. The decoder uses this encoded vector as input to generate the output sequence. Figure 2.4 illustrates this architecture. The encoder block and the decoder block are built using several recurrent units. These recurrent units can, for instance, be LSTMs, which were covered in Section 2.2.3.

2.2.5 The Attention Mechanism

The hidden vector, also denoted the context vector, between the encoder and the decoder blocks in Seq-2-sec models from Section 2.2.4, was discovered to be a limitation in Encoder-Decoder architectures. This limitation motivated the invention of the attention mechanism. Attention extends the Encoder-Decoder by passing all the hidden states from the encoder block to the encoder (Bahdanau et al., 2014). The decoder can then

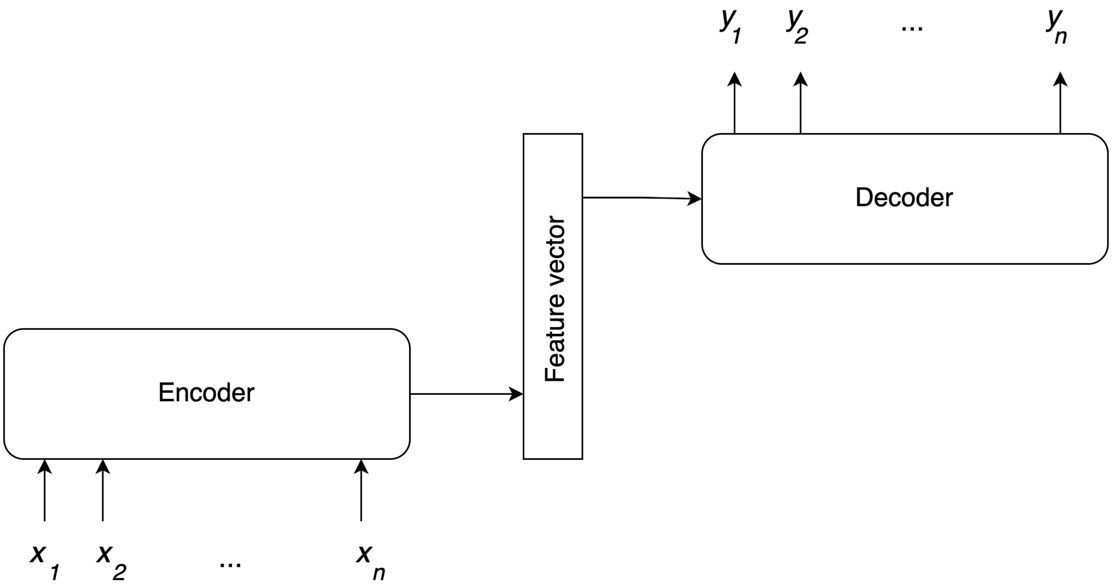


Figure 2.4: An illustration of the Encoder-Decoder architecture.

examine all the hidden vectors, score them according to their relevance and pay attention to relevant parts when processing a sequence.

2.2.6 The Transformer Architecture

Transformers use the attention mechanism from Section 2.2.5 and were first introduced by Vaswani et al. (2017). In short terms, the transformer is a Sequence-to-sequence architecture, consisting of a stack of encoders and a stack of decoders. Each encoder consists of a self-attention layer and a feed-forward network. The self-attention lets the model look at other positions in a sentence when encoding each word. The decoder blocks have a self-attention layer, followed by an encoder-decoder network, and lastly, a feed-forward network.

A new era within natural language processing started with the release of the transformer architecture. The architecture relies solely on the use of attention, and there is no recurrence used. Still, transformers have revolutionized the field of natural language processing. Since the Transforms uses attention rather than recurrence, parallelization is also more feasible, which is another advantage.

A wide range of pretrained language models using attention, transformers, has been released since Vaswani et al. (2017). The Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al., 2019) is a multilayer bidirectional transformer encoder based on Vaswani et al. (2017). BERT was pretrained the Wikipedia and Book Corpus and performed remarkably well on NLP tasks when released. Following is a further description of the transformers used in the experiment of this thesis.

Generative Pre-Traning (GPT) and GPT-2

Radford et al. (2018) proposed the first Generative Pre-Traning (GPT) model. They utilized a large corpus of unlabeled text data to generative pretrain a language model, which then can be finetuned for specific tasks. The GPT model uses the transformer architecture (Vaswani et al., 2017), is built using 12 layers of decoder-only transformers, and pretrained using the Book corpus. GPT was evaluated by Radford et al. (2018) and achieved state-of-the-art results on nine out of 12 NLP tasks tested.

Building upon GPT, the GPT-2 model was released by Radford et al. (2019). GPT-2 is based on the same architecture as GPT, but with increased vocabulary and context size. A new corpus, WebText, was gathered and used for the pretraining of GPT-2. The WebText corpus is collected by scraping data from 45 million web links, starting on Reddit and following high-quality links. GPT-2 was tested on eight tasks in a zero-shot manner, meaning the model was not finetuned for specific tasks upfront. Still, GPT-2 achieved state-of-the-art results on seven out of the eight tasks.

Enhanced Representation through Knowledge Integration (ERNIE) and ERNIE 2.0

ERNIE (Sun et al., 2019) is a language model inspired by the masking strategy used by the BERT model (Devlin et al., 2019). Besides the basic masking strategy from BERT (Devlin et al., 2019), two supplementary masking strategies are used by ERNIE to learn knowledge about phrases and entities in order to achieve better generalization and adaptability. These two strategies are respectively phrase-level strategy and entity-level strategy. Instead of masking only single words or characters, a phrase or an entity is treated as one unit and masked together during the training.

The ERNIE 2.0 model Sun et al. (2020) is based on the former ERNIE model described above, which hereinafter is denoted ERNIE 1.0 to distinguish between the two ERNIE models clearly. ERNIE 1.0 was specially tailored for the Chinese language, whereas ERNIE 2.0 is improved to perform better in the English language. ERNIE 2.0 is not only learning based on the co-occurrence of words but aims to capture lexical, syntactic, and semantic information from the training data.

The architecture of ERNIE 2.0 uses a multilayer transformer with encoders as proposed by Vaswani et al. (2017). The English ERNIE 2.0 is pretrained on data from Wikipedia, the Book corpus, data collected from Reddit, supplemented with the Discovery dataset (Sileo et al., 2019). For comparability with BERT, Sun et al. (2020) also use the same model settings as Devlin et al. (2019). The results reveal that the English base version of ERNIE 2.0 outperforms BERT on all ten tasks tested by Sun et al. (2020).

2.3 Fundamentals of Text Processing

Within text processing and text analytics, an instance of a text is often referred to as a document, and a collection of documents is a corpus. This section will introduce the fundamental basis of how text can be preprocessed and represented in meaningful ways.

2.3.1 Text Preprocessing

Operations can be applied to a document to prepare the text for further applications. *Segmentation* is the process of separating a text into sentences, and *tokenization* split each sentence into single tokens. A token is the most minor, meaningful semantic unit of the document. For example, words and numbers are tokens that together can make up a meaningful sentence.

Stemming and *lemmatization* are frequently used for text normalization and can be applied to normalize the text after a document is split into tokens. The purpose of stemming is to remove affixes of words by using rules for slicing the words. By using stemming, both "computer" and "computers" are reduced to "computer". Lemmatization interchange words with their lemma, the headword of a word which would be looked up in a dictionary. "algorithms" and "algorithmic" will both be interchanged with "algorithm".

Stopword removal can be done to reduce the corpus size and to increase the proportion of meaningful words. Stopwords are words that are frequently used in texts, for instance: "a", "the", "for", and "is". These words frequently appear in texts and thus have a minor discriminatory effect when analyzing documents.

2.3.2 Text Representation

Transforming text to representations is necessary for most text analytics applications and allows for more advanced processing. This section will cover the central methods for text representation.

Vector Representation

The fundamentals of the methods which will be covered are founded based on vector representations. A corpus' vocabulary is all the terms that exist in the documents that make up the corpus. The basis for vector representations is then a vector whose length equals the size of the vocabulary. For each document in the corpus, a vector on this form can represent the document. Each document encodes as a vector, and a position in the vector represents a given term. The specific model used decides how to compute each element in the vector.

Bag-of-Words Encoding

Bag-of-Words models encode text without preserving the order or relation of words. They simply tell which words are present in a document. One-Hot encoding is a boolean vector representation where the vector tells whether a term is contained in a document or not. Frequency-based encoding can similarly encode the document by counting the number of times a term appears in a document.

n-grams

n-grams is a technique for text representation that, to some degree, can preserve word order. *n*-grams are constructed by sliding a window of size n over the text and identify all subsequences. When $n = 1$, only single words will be included (unigrams). Bigrams (*n*-grams with $n = 2$) handles tuples of words. Trigrams (*n*-grams with $n = 3$) work on tuples of length three and following for greater sizes of n.

2.4 Natural Language Processing

The field of natural language processing (NLP) unites linguistics, computer science, and artificial intelligence (Chowdhary, 2020). Languages are for communication, and making it possible for computers to process natural language enhances numerous applications. This section will cover some of the techniques and subjects that are used when processing natural language. The subfield of NLP concerning text generation, natural language generation (NLG), is also included.

2.4.1 Language Modelling

Language modeling is the task of building models for predicting the next word given the previous words or the surrounding words. *Causal language modeling* concerns predicting the next token following a sequence of tokens. Hence causal models look only to the left side of the input token. Models using *masked language modeling*, on the other hand, receive an input where some of the input words are interchanged with a masked token. Masked language models thus look at both left and right sides of the masked tokens and use the full context to predict which word is most probable and should replace the mask token

2.4.2 Natural Language Generation

The field of natural language generation (NLG) concerns producing natural language from non-linguistic input. NLG covers a wide range of tasks, from machine translation to text summarization and dialogue systems like chatbots. Text generation can be divided into three subfields: data-to-text, text-to-text and image-to-text.

Data-to-text means generating natural language given input data fields. To illustrate, given data points of the current temperature outside, what time it is, and whether it is raining or not. A data-to-text system could generate human-like weather forecasts based on the data points. Automatic text summarization is an example of text-to-text NLG, where a system is given longer texts and reduces them to a shorter summary. Image-to-text also denoted as image captioning, generates text based on images.

2.4.3 Evaluation of Natural Language Generation

The field of natural language generation is rising. However, the lack of efficient and suitable methods for evaluations of NLG tasks is a bottleneck (Sellam and Parikh, 2020).

In broad, two main methods for evaluating NLG exist, automatic evaluation metrics and human assessment.

Examples of methods for automatic evaluation include BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004). BLEU (Bilingual Evaluation Understudy) is a metric for evaluating machine translation quality, whereas ROUGE (Recall-Oriented Understudy for Gisting Evaluation) can be used to evaluate both machine translation and automatic summarization. BLEU and ROUGE are the most utilized automatic metrics within the field (Sai et al., 2020).

Human evaluation can be done by creating a questionnaire and asking humans to rate generated text according to given criteria. Such evaluation can require extensive setup and be time-consuming, depending on the scope. Best practices for the human evaluation of automatically generated texts say always conducting a human evaluation when possible and use guidelines for designing the assessment, doing the measurement, and reporting the results (Van Der Lee et al., 2019).

2.4.4 Tools for Natural Language Processing

This section will present two tools within the natural language processing domain that are relevant for this thesis. First, will Hugging Face be introduced, a tool used in this thesis to utilize state-of-the-art language models. Secondly, the Linguistic Inquiry and Word Count (LIWC) program is explained. LIWC is a program frequently used in the literature to analyze text.

Hugging Face

The Transformers library (Wolf et al., 2020) released by Hugging Face is an open-source library for natural language processing, providing seamless access to and use of state-of-the-art language models. Besides providing easy access to the models themselves, utilities for data preparation, tokenization, and training are also given. As mentioned, the library is open-source, which allows the community to contribute by uploading new transformer models.

Linguistic Inquiry and Word Count

Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2015) is a language analysis program used to analyze text by assigning categories to the words. LIWC has predefined more than 70 classes and recognizes which words in a text belong to which classes. Examples of classes defined can be *Negative emotions* and *Positive emotions*. The program can calculate the percentage distribution of words from different categories, which can be helpful, for example, for determining whether a text consists of negative emotions.

3 Automatic Personality Prediction and Author Profiling

A structured literature review on automatic personality prediction from social media data was carried out as part of the specialization project preparing for this Master's Thesis. The research goal of the specialization project was to explore the field of automatic personality prediction from social media data, and a literature review was necessarily carried out. The structured literature review protocol detailing the review procedure can be found in Appendix C. This chapter will summarize the findings from the literature review that is considered relevant for this thesis.

The knowledge obtained from the structured literature review in the specialization project is highly relevant for this thesis for at least three reasons. First, it establishes which personality model should be applied. Secondly, datasets of social media text labeled with personality traits were identified. Lastly and most importantly, it is crucial to know the characteristics of expected writing styles concerning different personality traits, ages, and genders when evaluating personalized text from natural language generation systems. With this knowledge, the generated texts can be assessed against to what degree expected characteristics are present. Hence a presentation of characteristics of writing style with respect to personality, age, and gender are given in this chapter.

3.1 Modelling of Personality

The structured literature review of the specialization project found the Big Five personality model established as the most popular within the field. Other personality models were occasionally mentioned, but the Big Five model was dominating. The Big Five model is also argued to be the most researched personality model (Golbeck et al., 2011b; Kumar and Gavrilova, 2019), uniting the field of personality psychology into one personality model (Bachrach et al., 2012). Due to the establishment of the Big Five personality model as the leading model both in personality psychology and automatic personality prediction, it is unquestionably considered a suitable choice of personality model in this thesis.

3.2 Data Extraction and Datasets

Researchers within automatic personality prediction have used both existing datasets and manually collected and annotated their own datasets for automatic personality prediction from Facebook and Twitter. However, two published datasets stand out, the myPersonality dataset of Facebook data and the PAN15 Author Profiling dataset of Twitter data.

The myPersonality dataset was collected through a Facebook application where users voluntarily took a personality questionnaire and measured their Big Five personality traits. The scores were collected and used to build a dataset with Facebook profiles from 2.4 million users and their corresponding score on the Big Five personality traits. The myPersonality dataset had a significant impact by providing researchers with an enormous annotated dataset and undoubtedly facilitated research within the field of automatic personality prediction from Facebook data. As a part of the Workshop on Computational Personality Recognition 2013, a subset of the myPersonality data set was provided (Celli et al., 2013). The data set for the workshop used 250 users and contained their Facebook statuses, personality labels, and social network features.

The PAN15 Author Profiling (Rangel et al., 2015) dataset was published for the PAN 2015 Author Profiling Task and consists of tweets in English, Spanish, Italian, and Dutch. Remarkably, the PAN15 Author Profiling dataset was the only published dataset labeled with Big Five personality scores that were discovered by the literature review of the specialization project.

3.3 Feature Engineering for Automatic Personality Prediction

For automatic personality prediction, various features and text representations are utilized. The features extracted can be divided into linguistic-based features from the written texts and features representing metadata and users' profile information. The linguistic-based features are the only ones considered relevant to this thesis, and hence only those will be described in this section.

A finding from the structured literature review of the specialization project is that the Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2015) tool is commonly used to analyze the written language and produce textual features. The tool analyzes text according to predefined categories and counts each category's relative occurrence. An introduction to LIWC was also given in Section 2.4.4. The LIWC tool relies on predefined knowledge (Schwartz et al., 2013), thus using a closed vocabulary approach (Park et al., 2015).

An open vocabulary approach, on the other hand, does not require predefined categories upfront. Bi-grams is an example of an open vocabulary method, requiring no predefined categories. Extracting words, phrases, and topics in an open vocabulary manner is found to perform better than LIWC for predicting personality traits (Schwartz et al., 2013). Findings indicate that open vocabulary methods can discover new insight in correlations between language and author attributes.

Especially when doing feature engineering of tweets, it should be noted that it is common to analyze the number of retweets, mentions, URLs, and hashtags (Golbeck et al., 2011a; Preotiuc-Pietro et al., 2016).

3.4 Algorithms for Automatic Personality Prediction

The problem of automatic personality prediction can be modeled two in different ways that influence available algorithms. The problem can be tackled as either a classification problem or a regression problem. Automatic personality prediction as a classification problem will try to solve whether a user is entirely introverted or extroverted. On the other hand, with automatic personality prediction as a regression problem, the task is to predict to which degree a user is extroverted and predicting a real-valued score on a scale.

For the classification problem, a wide range of well-known algorithms is tested, such as Support Vector Machines, Random Forest, and Gradient Boosting. However, there is no clear trend on any algorithms consistently performing better than others. The same applies to the regression problem where various suitable algorithms are utilized, but no method stands out as more successful than others.

3.5 Statistical Analysis

The structured literature review discovered that extensive work analyzing correlations between personality traits and linguistic features is conducted in the field. Some of the findings include that use of articles (*a*, *an*, and *the*) correlate with older males and persons scoring high on openness (Schwartz et al., 2013). The LIWC category Anger is predictive for users scoring low on agreeableness and consciousness and for users scoring high on neuroticism. (Schwartz et al., 2013). The word you and the use of positive emotional words were also more often used by people scoring high on agreeableness (Golbeck et al., 2011a,b). On the other hand, swear words, words related to death, and negative emotions are negatively correlated with conscientiousness (Golbeck et al., 2011a).

3.6 Representation of Real-Life Personality on Social Media

The question of whether personality exposed on social media reflects users' actual personality or an idealized version is raised in the literature (Golbeck et al., 2011a,b; Kumar and Gavrilova, 2019). If social media users create idealized digital representations of themselves, inferring personality from social media can be misleading and not represent the real-life personalities (Carducci et al., 2018). Fortunately, Facebook profiles are shown to reflect actual personalities, and users are found not to decorate an idealized version of themself on Facebook (Back et al., 2010). This is supported by the finding that humans can predict others' personality traits based on their Facebook profiles, which would not be possible if the real-life personality was not exposed on Facebook (Bachrach et al., 2012).

3.7 Author Profiling

Author profiling concerns the task of identifying characteristics of authors based on text they have written. Aspects of automatic prediction based on personality are covered in the previous sections, so this part will concern author profiling in terms of predicting the age and gender of authors. Author profiling is a widely researched domain, and only a brief overview relevant to the thesis will be covered.

Argamon et al. (2003) explored the differences in male and female texts. Females were found to write more involved than males and to use more pronouns and negations in their writings. The males, on the other hand, used more determiners, quantifiers, and prepositions.

As a part of their Master's thesis, Berg and Gopinathan (2017) analyzed differences between social media texts written by males and females. They found females to use the heart emoticon (*<3*) three times more often than males. Regards emoticons, such as *:)* and *:-)*, females used more emoticons without a hyphen (:)) whereas males used more with a hyphen (*:-)*).

Schler et al. (2006) analyzed differences in writing styles based on a corpus of blogs. Concerning the differences between ages, they found that with the increased age of the author, the language also evolved. Pronouns, prepositions, and determines were used more frequently within older ages. Besides correlations between language and personality, Schwartz et al. (2013) also examined the effect of age on the language in social media texts and found that younger people used more emoticons than the elder.

4 Related Work

This chapter will describe how a structured literature review covering the state-of-the-art within personalized natural language generation is carried out and present the findings. The first section details the process of the structured literature review. The following sections start by distinguishing between subtasks of personalized natural language identified from the literature review. Then follows a presentation of datasets and models used in the literature, and the identified evaluation procedures within personalized natural language generation are covered. The chapter will end with examining the findings from the literature review in light of implications and motivation for the rest of the thesis.

4.1 Structured Literature Review

A structured literature review is conducted to gain sufficient knowledge within the field of personalized natural language generation. The method used for the literature review is based on Kofod-Petersen (2018). The motivation of using a structured literature review is for the author to gain an unbiased understanding of the field and enable reproducibility, as all steps of the process are documented in a review protocol. The full review protocol can be found in Appendix A. The structured literature review was carried out in three steps; planning, conducting and reporting.

4.1.1 Planning the Structured Literature Review

As a part of the Master's Thesis, there was a need for a structured literature review to answer Research Question 2 and gain necessary insight in related work for Research Question 3.

Research Question 2 *What are suitable methods for generating personalized natural language?*

Research Question 3 *What are suitable and efficient methods for evaluating personalized natural language generation systems?*

Following the methodology of structured literature review, a review protocol was developed and iteratively adjusted when necessary. The protocol can be found in Appendix A.

Table 4.1: The search terms for the structured literature review on personalized natural language generation.

	Group 1	**Group 2**
Term 1	NLG	Personalize
Term 2	Natural language generation	Customize
Term 3	Text generation	Personality

4.1.2 Conducting the Structured Literature Review

The process of carrying out a literature review can be formulated in five steps, as described by Kofod-Petersen (2018).

Step 1: Identification of Research

The first step when conducting the review was to decide upon the search domain and search terms. Google Scholar was considered the right choice of search domain because of its ability to find research from multiple academic resources and its built-in ranking process. The chosen search terms can be found in Table 4.1. Group 1 of search terms was included to obtain research within natural language generation in general. The terms in Group 2 aimed to target the personalization aspect of text generation. When concatenating the terms according to the groups, the search string follows as:

```
(NLG OR Natural language generation OR Text generation) AND
(Personalize OR Customize OR Personality)
```

The results from the search gave a total of 23 700 papers. Some adjustments to the terms and the search string were tested. See the details in Appendix A. Nevertheless, it was decided to keep the proposed terms and search string. The first 70 papers ranked by Google Scholar were collected for the next steps in the structured literature review. This was done to be realistic with the scope and due to observation of decreased relevance beyond the first 70 papers.

Step 2: Selection of Primary Studies

A selection of primary studies from the 70 papers extracted must be made. Primary inclusion criteria and secondary inclusion criteria were defined; both can be found in Appendix A. The selection of primary studies was performed in a two-step process. All papers were first assessed against the primary inclusion criteria. The papers passing the assessment were then evaluated against the secondary inclusion criteria. After the two steps, the remaining set of papers was reduced to 14 papers.

Step 3: Quality Assessment of Studies

The quality criteria can be found in Appendix A and are solely chosen as Kofod-Petersen (2018) provided. Each paper was assessed against and scored for all quality criteria. If it was fully fulfilled, 1 point was given, 1/2 point if it was partly fulfilled, and corresponding 0 points if it was not met. All the 14 papers selected in the previous step passed the quality assessment by obtaining a high score in total.

Step 4: Data Extraction and Monitoring

Data fields to be extracted from the primary studies were defined, see Appendix A.

Step 5: Data Synthesis

For all primary studies, data were collected and are provided in Table 5.2.

4.1.3 Reporting the Structured Literature Review

Table 4.2 reports the results of the structured literature review with respect to the chosen data fields to be extracted. This section gives a short synopsis of the findings before the following sections will give a more detailed presentation concerning the task, datasets, and models used, and how the results are evaluated.

First and foremost, it can be seen that the most common tasks within the papers are to either generate emotional coherent and polite texts (IDs 1, 2, 5, and 6) or to generate coherent and relevant dialogue responses (ID 3, 4, 7, and 13). One paper combines NLG with images and aims to generate image captions (ID 8). ID 9 and 10 examine how to control the style of generated texts, whereas ID 14 aims to generate personalized recommendations. Note that ID 11 is the only one that uses a personality model explicitly to generate personalized natural language.

Regards architectures and models, all research papers use sorts of deep learning models. The most popular is the Seq-2-seq architecture used in five of the papers. Transformers are only used in three of the papers, whereas none of these three uses the exact same transformer. No single dataset or data source stands out as most commonly used. Five papers use only self-collected data. Seven papers use existing datasets, and two papers combine existing datasets with collecting their own data.

Table 4.2: The extracted data fields from the identified literature from the structured literature review on personalized natural language generation.

ID	Author(s)	Title	Year	Task description	Models	Data set	Relevant findings and conclusions
1	Sun, Peng & Ding	Emotional Human-Machine Conversation Generation Based on Long Short-Term Memory	2018	Generate emotion-consistent responses to a post.	LSTM with an Encoder-Decoder framework.	Weibo posts and replies/ comments, made available for NLPCC 2017.	Slightly better results than related work in terms of emotion consistency.
2	Niu & Bansal	Polite Dialogue Generation Without Parallel Data	2018	Generate polite responses that are contextually relevant.	Three proposed models: a Fusion model, a label-fine-tuning model, and a reinforcement learning model.	Stanford Politeness Corpus and *MovieTriples* dialogue corpus	The Fusion model achieves politeness with poorer context relevance. The two other models were able to produce significantly more polite responses without sacrificing dialogue quality.
3	Herzig, Shmueli-Scheuer, Sandbank & Konopnicki	Neural Response Generation for Customer Service based on Personality Traits	2017	Generate customer service responses conditioned on a target personality.	Seq-2-seq architecture with a layer representing personality and a hidden layer for learning high-level personality-based features.	A dataset of 1 million customer service conversations.	Results outperform baseline Seq-2-seq model on BLEU scores.

Continued on the next page

Continued from previous page

ID	Author(s)	Title	Year	Task description	Models	Data set	Relevant findings and conclusions
4	Zhang, Zhu, Wang, Zhao & Liu	Neural Personalized Response Generation as Domain Adaptation	2019	Generate personalized responses in a two-phase approach.	RNN based Seq-2-seq model.	Self-collected.	The proposed model outperforms the state-of-the-art on language model personalization.
5	Zhou, Huang, Zhang, Zhu & Liu	Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory	2018	Given a post and an emotion category, generate a response that is coherent with the emotion category.	Seq-2-seq architecture implemented with GRUs.	Use of NLPCC2013 and NLPCC2014 datasets and STC dataset.	Able to generate responses that are coherent in both content and emotion.
6	Ghosh, Chollet, Laksana, Morency & Scherer	Affect-LM: A Neural Language Model for Customizable Affective Text Generation	2017	Generate affective sentences for a target emotion with varying degrees of affect strength.	LSTM with a term to represent affective information.	Fisher English Training Speech Corpus, Distress Assessment Interview Corpus, SEMAINE dataset, Multimodal Opinion-level Sentiment Intensity Dataset.	The proposed model generates naturally looking emotional sentences without sacrificing grammatical correctness.

Continued on the next page

Continued from previous page

ID	Author(s)	Title	Year	Task description	Models	Data set	Relevant findings and conclusions
7	Zhang, Sun, Galley, Chen, Brockett, X. Gao, J. Gao, Liu & Dolan	DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation	2020	Generate relevant, contentful, and context-consistent conversation responses.	Proposes DialoGPT, an extension to GPT-2.	Collected data from Reddit, tested on DSTC-7 dataset.	Both human and automatic evaluation metrics show that the proposed model performs close to humans in generating conversational responses.
8	Shuster, Humeau, Hu, Bordes & Weston	Engaging Image Captioning via Personality	2019	Generate image captions with a personality to engage humans.	Built TransResNet using ResNet152, Transformers, and Feed Forward Neural Networks.	Collected a dataset, Personality-Captions.	The proposed model is shown to produce image captions close to matching human performance in terms of engagement and relevance.
9	Oraby, Reed, Tandon, Sharath, Lukin & Walker	Controlling Personality-Based Stylistic Variation with Neural Natural Language Generators	2018	Explore explicit stylistic supervision to neural networks to control style.	Seq-2-seq TGen, a system based on Seq-2-Seq generation with attention.	Built a corpus using Personage.	The most explicit model is shown to achieve high fidelity to both semantics and stylistic goals.

Continued on the next page

Continued from previous page

ID	Author(s)	Title	Year	Task description	Models	Data set	Relevant findings and conclusions
10	Ficler & Goldberg	Controlling Linguistic Style Aspects in Neural Language Generation	2017	Generate natural language text that conforms to a set of content-based and stylistic properties.	LSTM-based language model.	Corpus collected from Rotten Tomatoes.	Shown to successfully generate coherent movie reviews corresponding to linguistic style and content.
11	Keh & Cheng	Myers-Briggs Personality Classification and Personality-Specific Language Generation Using Pre-trained Language Models	2019	Explore the use of fine-tuned BERT model for personality-specific language generation.	BERT.	Self-collected from Personality Cafe.	BERT is better at generating language for extroverted personalities than introverted ones.
12	Golovanov, Kurbanov, Nikolenko, Truskovskyi, Tselousov & Wolf	Large-Scale Transfer Learning for Natural Language Generation	2019	Studies how pretrained language models can be applied and adapted for natural language generation.	OpenAI GPT.	PersonaChat dataset.	Results indicate that various architectures have different inductive biases regards the type of input context.

Continued on the next page

Continued from previous page

ID	Author(s)	Title	Year	Task description	Models	Data set	Relevant findings and conclusions
13	Qian, Huang, Zhao, Xu & Zhu	Assigning Personality/Profile to a Chatting Machine for Coherent Conversation Generation	2018	Generate chat responses that are coherent to a pre-specified personality or profile.	Encoder-Decoder architecture.	Self-collected from Weibo.	Model is shown to effectively generate responses that are coherent to pre-specified personality and profile.
14	H. Chen, X. Chen, Shi & Zhang	Generate Natural Language Explanations for Recommendation	2021	Generate free-text natural language explanations for personalized recommendations.	Hierarchical Seq-2-seq.	Amazon 5-core.	Improvement in recommendations accuracy and explanation quality.

4.2 Aspects of Personalization

Personalized natural language generation can be understood in several manners, as discovered during the structure literature review. This section aims to explain the various interpretations of personalization within natural language generation and describe how the identified literature targets and interprets the term. Keh et al. (2019) is the only paper from the structured literature review that bases personalization on a personality model to generate texts specific for given personality traits.

Sun et al. (2018), Niu and Bansal (2018), Zhou et al. (2018), and Ghosh et al. (2017) personalize and condition the text generation in terms of emotions and politeness. The task defined by Sun et al. (2018) is given a post with an assigned emotion category, generate a response that is coherent with the emotion category. That is similar to the problem formulated by Zhou et al. (2018), generating responses with different emotional states. Ghosh et al. (2017) investigated how to generate sentences with target emotions and to vary the degrees of emotions. Niu and Bansal (2018) targes similar problems in the manner of generating coherent responses, but having the model to generate polite answers to a response. Hence, personalization can thus be in terms of assigning emotions and politeness to generated texts.

Another interpretation of personalization and task description identified is the personalization of chatbots and dialogue systems. Herzig et al. (2017) modelled a customer chatbot to generate responses conditioned on a set of personality traits. Personalization of conversational systems was also targeted by Zhang et al. (2019), Zhang et al. (2020), and Qian et al. (2018). Lastly, personalization could also be to control the stylistic aspect of the text generated, tackled by Oraby et al. (2018) to vary the style of generated texts, but keeping the same message of the samples, and by Ficler and Goldberg (2017) to control the style of conditional movie review generations.

4.3 Datasets for Personalized Natural Language Generation

The results from the literature review show that using existing datasets is almost just as common as collecting own corpora. However, no main data source stands out as the most used in the research identified for either the existing or collected datasets. Recall that a detailed overview of which data used in the different papers is also found in Table 4.2.

The most established existing datasets used in identified literature are the *PersonaChat* (Zhang et al., 2018), *MovieTriples* (Serban et al., 2016), and *Amazon Review Data* (Ni et al., 2019) datasets. However, none of the datasets are used multiple times across the papers identified. Regards collecting own data, Qian et al. (2018) collected data from Weibo, Ficler and Goldberg (2017) from Rotten Tomatoes and Keh et al. (2019) from Personality Cafe. Both Zhang et al. (2020) and Shuster et al. (2019) used existing datasets supplemented with self-collected data.

4.4 Architectures and Models for Personalized Natural Language Generation

This section is dedicated to draw the lines and provide an overview of the different architectures and models identified in the structured literature review. First and foremost, all papers use some variants of deep learning techniques. The Sequence-to-sequence (Seq-2-seq) architecture, described in Section 2.2.4, stands out as the most utilized in the identified research as Herzig et al. (2017), Zhang et al. (2019), Zhou et al. (2018), Oraby et al. (2018), and Chen et al. (2021) report using the architecture. Moreover, the Long Short-Term Memory architecture, described in Section 2.2.3, is also used by Sun et al. (2018), Ghosh et al. (2017), and Ficler and Goldberg (2017).

Even though the structured literature review limited the search to papers published in 2017 and sooner due to that was when the attention mechanism was introduced by Vaswani et al. (2017), only three of the papers identified use transformers in their research. Golovanov et al. (2019), as the first, using GPT (Radford et al., 2018) to study how pretrained language models could be applied and adapted for natural language generation. Zhang et al. (2020) built their system around GPT-2 (Radford et al., 2019), and BERT (Devlin et al., 2019) was used by Keh et al. (2019).

4.5 Evaluation of Personalized Natural Language Generation

Findings from the literature review concerning the evaluation of personalized natural language substantiate that it is challenging. As Zhang et al. (2019) state, automatic evaluation is still an open problem. As seen in Section 4.2, multiple approaches and subtasks within personalized natural language generation are tackled, but the lack of automatic measures seem to be a problematic issue. However, a finding from the identified literature is that it is common to combine automatic metrics and human evaluation when possible.

Before heading over to manual evaluation of personalized generated texts, one automatic measure sticks out, namely the BLEU metric (Papineni et al., 2002) introduced in Section 2.4.3, which is used in the half of the papers identified (Herzig et al., 2017; Niu and Bansal, 2018; Oraby et al., 2018; Shuster et al., 2019; Golovanov et al., 2019; Zhang et al., 2020; Chen et al., 2021). However, BLEU is claimed to be unsuitable for a wide range of NLG tasks and is found not to correlate well with human judgments (Liu et al., 2017; Niu and Bansal, 2018; Zhang et al., 2019; Zhou et al., 2018) and hence not solve the issue of evaluating personalized natural language generation systems.

The use of manual assessment and human evaluation is widespread as well in the literature. Human evaluation can be used to measure whether generated samples are appropriate and coherent (Zhou et al., 2018; Sun et al., 2018). Criteria used for human evaluation of generated texts were relevance, informativeness, and human-likeness in Zhang et al. (2020). Qian et al. (2018) used criteria of naturalness (fluency and grammatical correctness), logic (whether the response is a logical reaction), and correctness (whether a response correctly answers a question) for human evaluation.

4.6 Implications and Motivation

The findings from the structured literature review on personalized natural language generation substantiate that there are significant gaps in the existing research. Fortunately, this is within areas this thesis aim to contribute. First and foremost, recall from Section 3.1 that the Big Five personality model was the preferred choice within automatic personality prediction. However, as discovered in this structured literature review, the Big Five personality model was never mentioned, and only one paper conditioned the personalized text generation on personality traits directly. Henceforth, this motivates exploring the use of the Big Five personality model for personalized natural language generation.

Moving on, no data sources are established as the better choice for personalized natural language generation. As presented in Section 3.2, social media text datasets labeled with Big Five personality traits exist and can be a proper fit if using the Big Five personality model as the psychological basis for personalized text generation.

Despite the recent success of the transformer architecture within natural language processing, remarkably few papers identified in the structured literature review utilized these models for personalized natural language generation. Personalized natural language generation using transformers seems thus not fully explored and should be further addressed.

5 Datasets

The following chapter will give a detailed presentation of the two datasets that will be used in the upcoming experiments. As introduced in Section 3.2, labeled datasets are made available for research. Two different datasets will be used in this thesis and are presented in this chapter. Section 5.1 explains the myPersonality dataset consisting of Facebook status updates. In Section 5.2, the PAN15 Author Profiling dataset consisting of tweets is presented.

5.1 myPersonality

The myPersonality-dataset was released for the Workshop on Computational Personality Recognition (Shared Task) 2013 and consists of Facebook status updates annotated with the Big 5 Personality scores. Recall from Section 3.2 that the data was gathered through a Facebook application that let users take a NEO PI-R test and give consent to allow their data to be used for research purposes. NEO PI-R was described in Section 2.1 and is an inventory for measuring the Big Five personality traits.

The myPersonality-dataset consists of 9 917 Facebook statuses from 250 different users. One row represents a single status update and will be denoted as a document. Each document in the dataset is labeled with the author's score on each personality trait in the Big Five personality model. Additional data fields include network size, density, brokerage, and transitivity. The personality scores are given in both numerical values and discrete classes. Each trait is scored as a real number between 0 and 5 to indicate to which degree the personality score is present for the user. This is the numerical value. And as mentioned, for each trait, a categorical label (*yes* or *no*) is also given and answering binary whether the trait is present or not.

The average length of the documents is 80.6 characters. The shortest statuses are only two characters long and is the following:

- *<3*,
- *:(*,
- *):* and
- *no*.

Table 5.1: Example row with relevant data fields from the myPersonality dataset.

Field	Value
#AUTHID	b7b7764cfa1c523e4e93ab2a79a946c4
STATUS	*is too lazy to put her stuff back in order. Maybe tomorrow.*
sEXT	2.65
sNEU	3.0
sAGR	3.15
sCON	3.25
sOPN	4.4
cEXT	n
cNEU	y
cAGR	n
cCON	n
cOPN	y
DATE	01/18/10 02:35 AM

The longest document has a length of 435 characters and illustrates the diverseness of the dataset when compared to the shortest documents:

> *Heh...:"God I wish that I could hide away//And find a wall to bang my brains//I'm living in a fantasy,//a nightmare dream...reality//People ride about all day//In metal boxes made away//I wish that they would drop the bomb//And kill these cunts//that don't belong! I hate people!//I hate the human race//I hate people!//I hate your ugly face//I hate people!//I hate your fucking mess//I hate people!//They hate me!"-Anti-Nowhere League.*

All proper names of persons in the documents are replaced with a fixed string (Celli et al., 2013). To exemplify, in the following status the specific proper name is interchanged with a **PROPNAME** tag:

> *happy birthday *PROPNAME*! Mommy loves you veryyyy much<3.*

An example row from the dataset is shown in Table 5.1. *#AUTHID* is a unique, anonymized identifier for each user in the dataset. *STATUS* is the raw Facebook status written by the user. *sEXT*, *sNEU*, *sAGR*, *sCON*, and *sOPN* are the numerical scores from 0 to 5 on each Big Five personality trait. Corresponding are the following attributes *cEXT*, *cNEU*, *cAGR*, *cCON*, and *cOPN* the binary values for whether the user is defined as belonging to the class for each trait, as described earlier in the section.

Additional data fields in the myPersonality-dataset include data about the Facebook network of the user. For example, the size of the network, the density, brokerage, and the transitivity in the network. In line with the research goal, data fields about the user's network will not be utilized in the research and are for simplicity left out of the description. See therefore Celli et al. (2013) for a further description of all data fields.

5.2 PAN15 Author Profiling

For the PAN (an organization that organizes series of events and shared tasks within authorship analysis and plagiarism detection) at the Conference and Labs of the Evaluation Forum (CLEF) 2015, a Twitter dataset was released for a shared task on author profiling. The dataset will be referred to as the PAN15 Author Profiling dataset (Rangel et al., 2015).

The dataset consists of several tweets written by anonymized users. The tweets are annotated with the user's age group and scores on Big Five personality traits. Scores on the personality traits are obtained by having the users take the Big Five Inventory 10 (BFI-10, see also Section 2.1) test.

The original dataset for the PAN15 Author Profiling task consists of tweets in both English, Spanish, Italian and Dutch. Only English tweets are relevant for this project, so for simplicity, only the English part of the dataset will be described and taken into use. The original dataset is partitioned into a training part and a test part. Labels are provided for both, so the two parts will be concatenated and handled as one dataset in this thesis. The merge between the training and test part is done to make available as much data as possible for the experiment. Also, the experiment will not require a training and testing split, so it is not necessary to keep the separation.

The PAN15 Author Profiling dataset consists of 27 344 documents from a total of 294 different users. A great majority (80%) of the users have written more than 90 tweets each. The average length of tweets in the dataset are 77.3 characters. The shortest tweet appearing in the dataset is only a single character, *o*, whereas the longest is 192 characters long:

> *????????? @username: WTF “@username: ????????????????? @username: ???????? “@username: ???????????????????????????????? ????????????????????????on to the next level ?????? @username: Seriously?.*

When seeing this example, note that data cleaning procedures will be examined in Section 7.2.2.

An example instance from the PAN15 Author Profiling dataset is shown in Table 5.2. The *userid* is the unique identifier for each user in the dataset. The *tweet* is the tweet posted on Twitter. From the example in Table 5.2, it can be seen that hashtags (in this case *#MeMyself&I*) are not replaced by a general tag. The same applies for URLs in tweets, which are kept and not replaced with general tags. Mentions of other Twitter users, on the other hand, are replaced with a standarized *@username* tag.

Table 5.2: Example row from the PAN15 Author Profiling dataset.

Field	**Value**
userid	bb88ec91-6085-4a89-94e0-6ae12a3afd3a
tweet	*It's all about me now. I gotta do what I gotta do to make it in this world & make sure that I'm okay. #MeMyself&I*
gender	F
age_group	18-24
extroverted	0.1
stable	0.2
agreeable	0.0
conscientious	0.3
open	0.3

The *gender* says whether the tweet is written by a male (*M*) or a female (*F*). The *age_group* is one of the following options:

- *18-24*,
- *25-34*,
- *35-49* or
- *50-XX*.

gender and *age_group* are self-reported by the user. The *extroverted*, *stable*, *agreeable*, *conscientious*, and *open* are Big Five scores normalized on a scale from -0.5 to +0.5.

6 Architecture

This chapter will describe the architecture implemented to realize the system for personalized natural language generation of social media short-text. The first section gives an overview of the complete architecture, before the following sections explain each of the parts which together compose the system.

6.1 An Overview of the Full Architecture

The system can be broken down into four parts, which are detailed in separate subsections below and illustrated in Figure 6.1. This section gives an overview of the architecture, explains the system flow and the relevance and importance of each components. Figure 6.1 shows how data is preprocessed and fed to models for finetuning. The finetuned models are used to generate results which are then evaluated.

First and foremost, the datasets presented in Chapter 5 must be preprocessed so that the data can be utilized. The raw data are provided in different formats and folder structures, hence preprocessing to bring the datasets together must be done. This process is the data preprocessing procedure which is detailed in Section 6.2.

The next part is the finetuning of the pretrained language models. The pretrained models GPT-2 and ERNIE 2.0 provide a basis, and the preprocessed data is used for finetuning, which is making adjustments to specialize the models for the task of personalized natural language generation. Section 6.3 explains this part of the architecture.

A central part of the architecture is the personalized text generator, which Section 6.4 describes. The architecture is designed to support two different methods for personalized text generation. Either by specifying gender, age, and selected personality traits of the author or by specifying the degree of each of the Big Five personality traits.

Lastly, the results are evaluated against defined evaluation procedures. Section 6.5 describes this part of the architecture and the automatic evaluation procedure.

6.2 Preprocessing of the Datasets

For personalized text generation based on social media data, the datasets presented in Chapter 5 must be made suitable for the task. The two datasets must be concatenated together and prepared for the finetuning process. This section will describe the steps conducted for preparing the datasets. The dataset from Section 5.1 will be denoted as the myPersonality, similarily the dataset from Section 5.2 will be denoted as the PAN15 dataset.

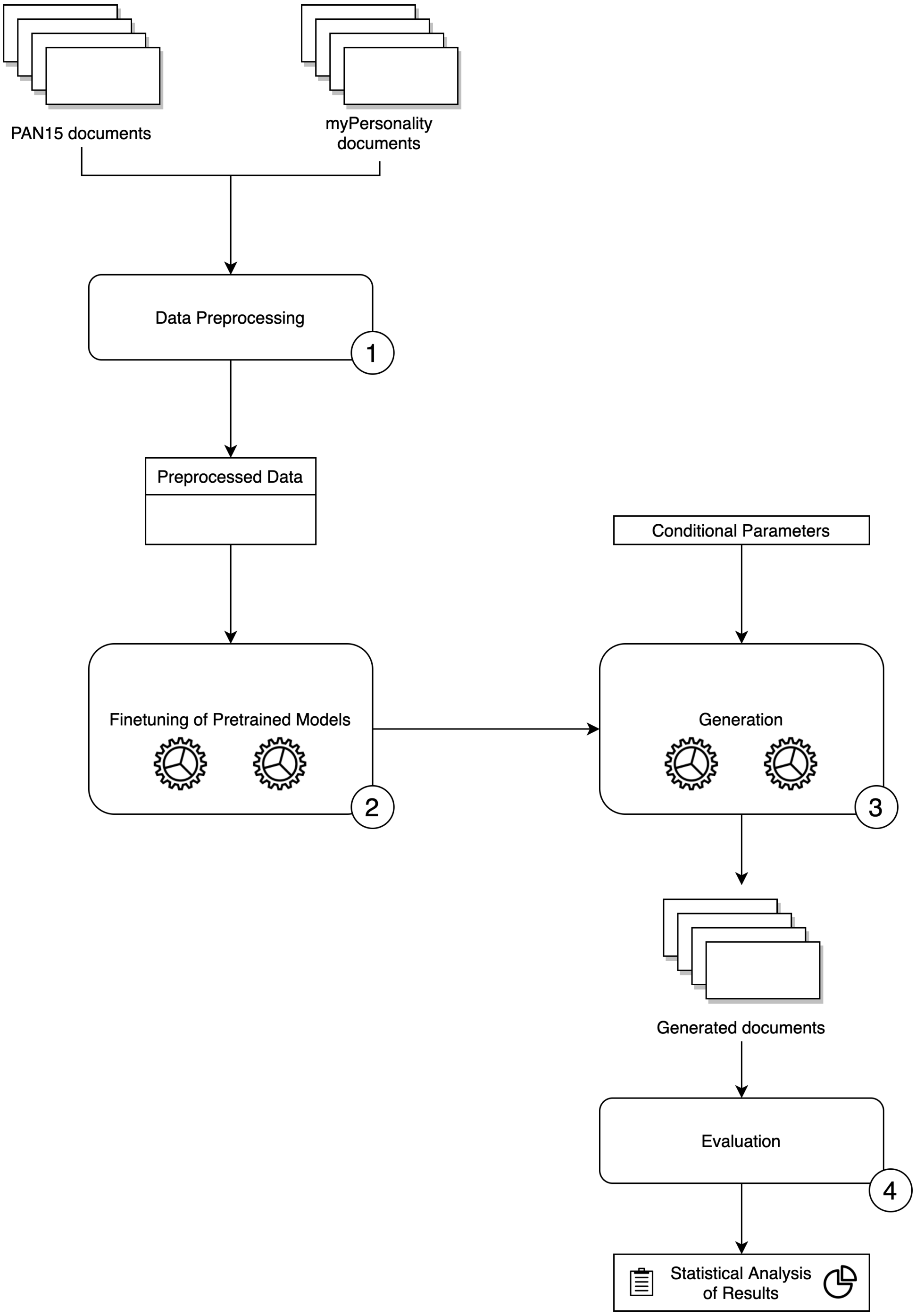


Figure 6.1: An illustration of the complete system architecture. First, data is preprocessed and then used for finetuning GPT-2 and ERNIE 2.0. The finetuned models are used to generate personalized natural language according to conditional input parameters. Lastly, the produced texts are evaluated.

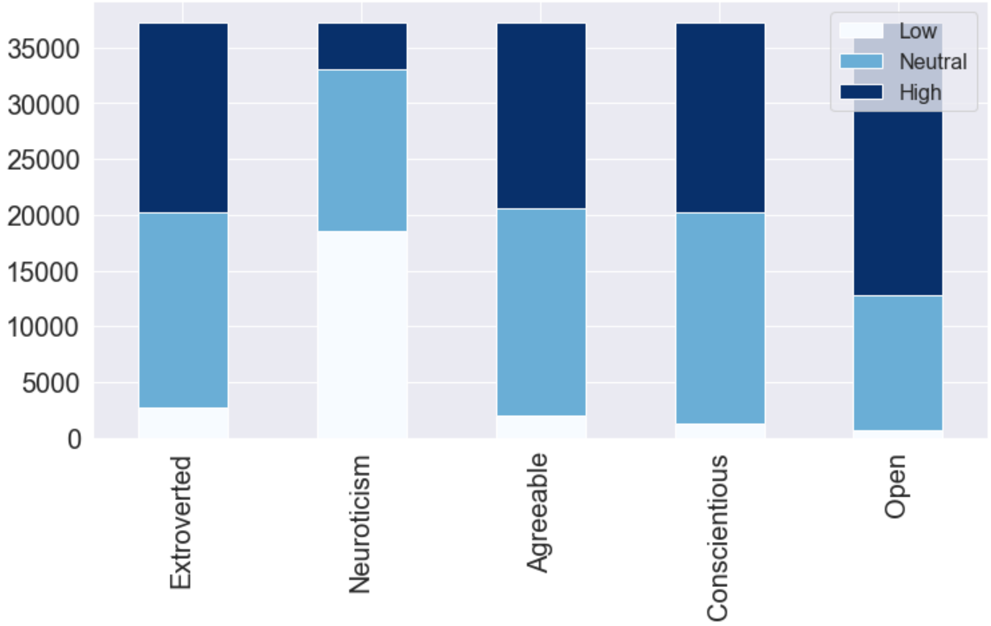


Figure 6.2: The distribution of classes per personality trait.

First and foremost, the datasets are provided in different formats. The myPersonality dataset is provided in a file of comma-separated values (.csv-file), making it very convenient to read and process. On the other hand, the PAN15 dataset is provided in an extensive folder structure of extensible markup language (.xml) files. Hence a more throughout process of reading and preparing the PAN15 data must be carried out.

Next, both datasets consist of data fields that are outside the scope of this project. Thus the next step of the processing is to remove irrelevant columns. The user identification, the written status or tweet, and the scores on each personality trait are the data fields that will be kept for both datasets. In accordance, the PAN15 dataset provides fields for the authors' age group and gender, which are kept.

Recall that the Big Five personality model consists of five overall personality traits: openness, consciousness, agreeableness, extroverted, and neuroticism. The PAN15 dataset differs from this model by having a score on a trait *stable* instead of neuroticism. Emotionally stable is interpreted as being the opposite of neuroticism. Thus the scores on the stable trait in the PAN15 dataset are inverted and then treated as scores on neuroticism. This conversion is considered necessary to have consistency in the experiments' data and be aligned with the Big Five personality model.

The next step in the data processing is to normalize the personality trait scores. For the myPersonality dataset, each trait's scores are given as numerical values between 0 and 5. In the PAN15 dataset, on the other hand, the values for the traits are given as values between -0.5 and 0.5. In preparation for treating the instances as one dataset, the scores must be transformed to the numerical same scale.

The PAN15 dataset is of greater size than the myPersonality dataset. Thus the scores from the myPersonality dataset are chosen to be converted since it requires the least number of operations. A MinMaxScaler[1] is used for transforming each numerical score in the myPersonality dataset to the range from -0.5 to 0.5, which is the scale where the PAN15 scores are located. The MaxMinScaler works by scaling each value independently according to the mean and standard deviation of the population to ensure a correct distribution among the new scale.

At this point, the dataset represents each of the five personality traits as scores between -0.5 and 0.5. The problem of personalized text generation in this project will require the personality modeled as whether the author is either low, neutral, or high on each personality trait. Thus it is necessary to discretize the continuous values of the numerical scores. A KBinsDiscretizer[2] is used to distribute the data into three bins using a uniform strategy where all bins are of the same width. Hence, it is not necessarily the same amount of instances in each class, which is not an expected outcome either. The distribution of the classes after the discretization can be seen in Figure 6.2.

6.3 Finetuning of the Models

The pretrained models used to realize the architecture are GPT-2 and ERNIE 2.0. This section will describe the finetuning process and explain relevant aspects for the finetuning. Recall from the structured literature review that transformers only was used in the three of the 14 papers identified. GPT (Radford et al., 2018) and GPT-2 (Radford et al., 2019) were two of these, therefore it was chosen to continue the research and use GPT-2 in this thesis as well. The last model discovered in the literature review was BERT (Devlin et al., 2019). To test a model with similarities to BERT but being more current released and shown promising on other natural language processing tasks, ERNIE 2.0 (Sun et al., 2020) was chosen as the second model in the architecture.

Recall that a language model, in general, is a model that looks at a sequence of words (a sentence) and predicts which word should follow. GPT-2 works by predicting one new token to follow in a sentence at each timestep, hence GPT-2 is autoregressive, which means that when the model outputs a token, it is added to the input and used in the prediction of the next token. ERNIE 2.0, on the other hand, is trained using a masking strategy described in Section 2.2.6.

Both GPT-2 and ERNIE 2.0 are pretrained on existing datasets. To specialize the models for the Twitter and Facebook domains and prepare for conditional text generation, finetuning must be done. Finetuning in this manner means that some of the models' layers are unfrozen (meaning their weights can be adjusted during finetuning). Data is then fed to the model to learn the context between the controllable attributes (in this case personality, age, and gender) and the text in the documents.

[1] MinMaxScaler, scikit-learn `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html`

[2] KBinsDiscretizer, scikit-learn `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.KBinsDiscretizer.html`

Table 6.1: The available keywords for personalized natural language generation using the keyword version.

Type of keyword	Available options
Gender	`male`, `female`
Age group	`young`, `younger adult`, `adult`, `senior`
Personality traits	`introverted`, `extroverted`, `neuroticism`, `stable`, `agreeable`, `hostile`, `conscientious`, `spontaneous`, `open`, `closed`

Both GPT-2 and ERNIE 2.0 are available through Hugging Face[3] (Wolf et al., 2020). Hugging Face is an open-source natural language processing library providing easy access and use of state-of-the-art language models through their Transformers library. The source code for ERNIE 2.0 (Sun et al., 2020) is orginially released on PaddlePaddle, a Chinese research and development platform for deep learning tasks. Fortunately, the ERNIE 2.0 model is converted to Hugging Face's format by Hu (2019), and the pre-trained model can thus be directly loaded. Because of its simplicity, both GPT-2 and ERNIE 2.0 are loaded using the Transformers library.

In general, when datasets are used for pretraining, the data must be prepared to provide proper input for the models to utilize. Tokenization is a vital part of natural language processing. To tokenize a text means to split documents into smaller parts, often words or characters. Since language models can expect different input of training data, it is necessary to use appropriate tokenizers. Fortunately, specialized tokenizers tailor-made to the specific language models are also provided via Hugging Face and are consequently used in the architecture because of their simplicity.

To summarize, the pretrained models are loaded using the Transformer library from Hugging Face. The preprocessed datasets from Chapter 5 are tokenized using tailored tokenizers, also provided via the Transformer library. For the models to capture the relationship between conditional parameters and the written text, the attributes for each instance is embedded with the text before it is fed to the model for finetuning. All available data is used for finetuning, and the models are saved to disk when the finetuning is complete so that they can be used for generation. The setup and parameters used in the finetuning will be described in Section 7.2.4.

6.4 Text Generation Using Finetuned Models

After the pretrained models are finetuned for the specific task, the next part of the architecture is the part responsible for conditional text generation. The generation part takes conditional variables as input and produces text expected by the models to be written by a person with the given attributes.

[3] Hugging Face, `https://huggingface.co/`

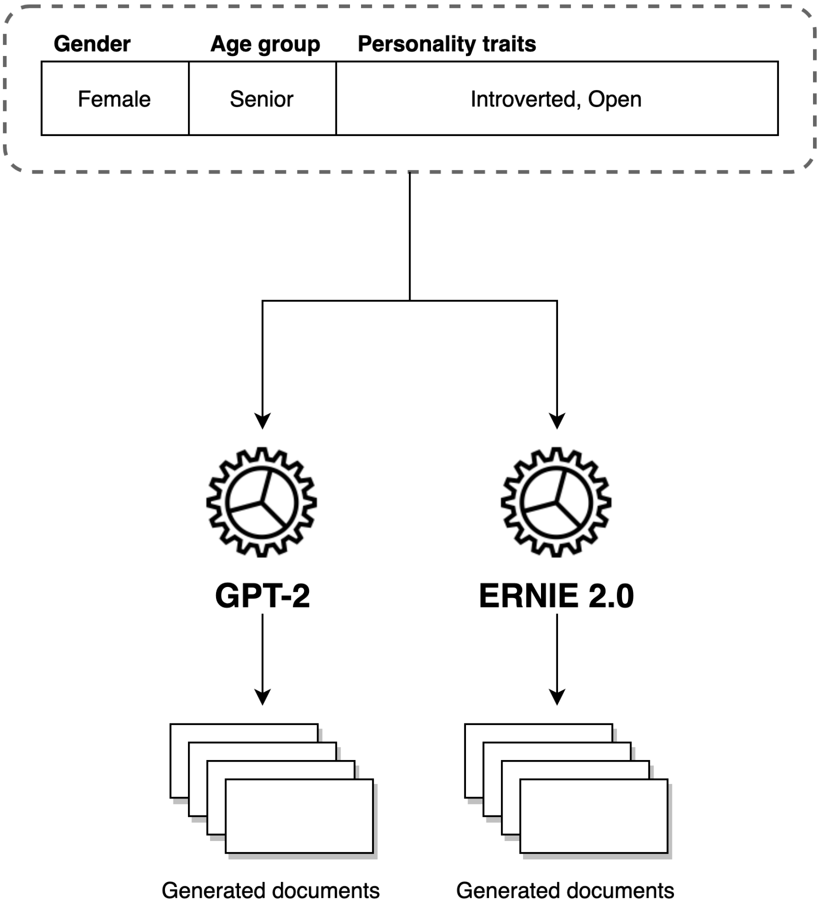


Figure 6.3: An example of supported conditional input for generation using the keyword input version. Keywords for gender, age group, and personality traits can be specified according to those provided in Table 6.1.

Recall that the architecture is designed to support two options for providing conditional parameters. The first option is to specify a fixed number of keywords describing the fictive author. Figure 6.3 illustrates this solution, and Table 6.1 the available keywords. The five personality traits of the Big Five model and the opposite traits are available keywords for controlling the personality. Note that it would be illogical to input opposite pairs (for example both *introverted* and *extroverted*) as those are antonyms and in direct conflict. The other available option in the architecture is to specify a *high*, *neutral*, or *low* score on each of the Big Five personality traits. Figure 6.4 presents this alternative. Note that only the keyword input supports specifying gender and age group in accordance to personality traits. Consequently, the keyword input format is only pretrained using the PAN15 dataset since age and gender data are only present in that dataset.

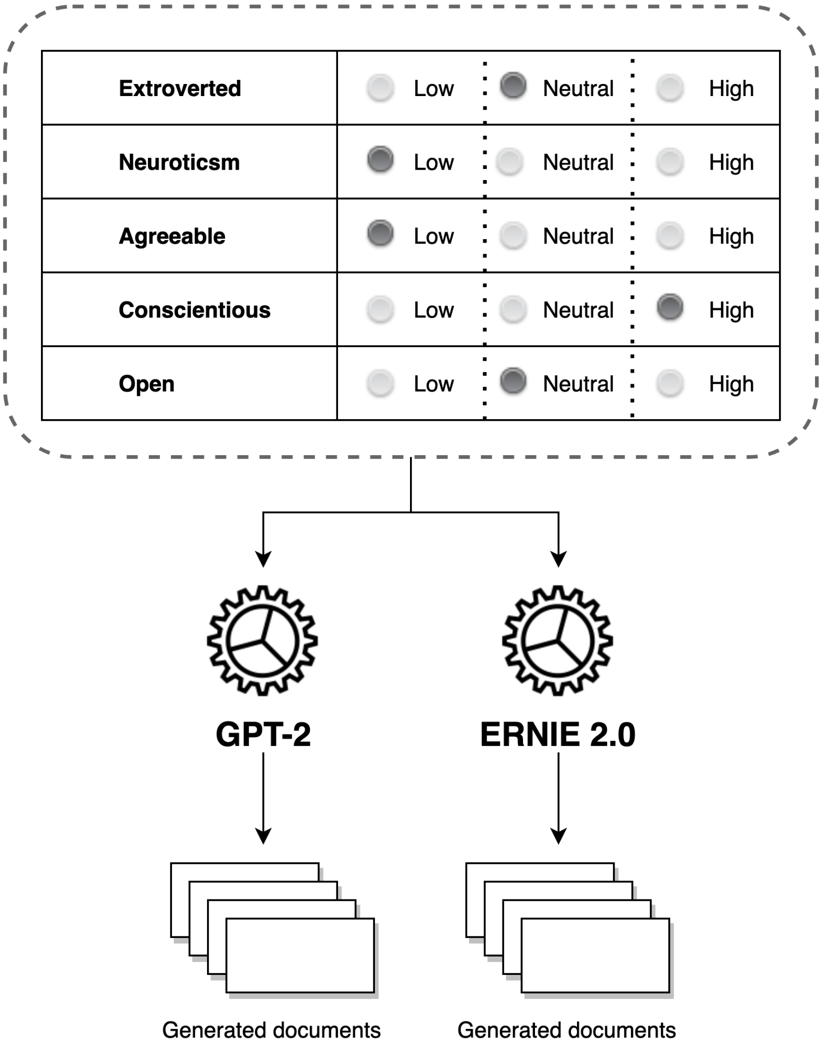
The generation takes place with a decoding strategy using a combination of Top-K and Top-p sampling. Top-K sampling is a strategy for generation where the K most likely words are extracted, and the mass probability is distributed among only those K words. Top-p sampling is another slightly different strategy. A minimum set of words whose probabilities sum to p is first chosen, and then the mass probability is distributed over those words. Top-K and Top-p are combined and used together in the architecture. The specific parameters to be used for generation will be covered in Section 7.2.5.

6.5 Evaluation of the Generated Texts

A vital aspect of the system is the component responsible for evaluating the generated text samples. Evaluating natural language generation systems is not a straightforward process since the produced samples cannot be assessed against a predefined solution since such a solution does not exist because of the task's nature of generating new texts. As discovered in the related work from Section 4.5, standard evaluation procedures for defined natural language generation tasks is an open issue and no standard metrics for evaluating personalized language are established.

Although no standard metric within personalized text generation is identified, that does not mean that the produced text cannot be analyzed and evaluated. The system is designed to support statistical and linguistical analysis of the produced text. Aligned with best practices for evaluating natural language generation systems, the produced text will also be evaluated by humans to quantify its grammatical correctness (Van Der Lee et al., 2019).

Recall from Section 3.5 that statistical analysis of tweets and Facebook status updates has shown correlations between personality and writing style. The system is designed to analyze the generated documents against these findings. To exemplify, older males are found to use more articles (*a*, *an*, and *the*). The generated documents are analyzed to evaluate whether articles are more present in the text conditioned on older males.



Extroverted	○ Low	● Neutral	○ High
Neuroticsm	● Low	○ Neutral	○ High
Agreeable	● Low	○ Neutral	○ High
Conscientious	○ Low	○ Neutral	● High
Open	○ Low	● Neutral	○ High

Figure 6.4: An example of supported conditional input for generation using the personality trait input version.

7 Experiments and Results

This chapter will present the experiments conducted. First, the chapter will provide the experimental plan created upfront, which will be continued with a description of the experimental setup. This description will include the parameters used in finetuning and generation, and the conditional input used to generate personalized natural language. Some initial experiments were conducted as a part of the setup to decide the optimal specifications, and a description of these will also be provided. Lastly, the experimental results of the generated texts will be presented.

7.1 Experimental Plan

An experimental plan is created to provide structure for the experiments and outline the procedure from setup to evaluation. The experiments to be conducted are in line with the research goal introduced in Section 1.2: *Contribute to the field of personalized natural language generation by exploring methods for the generation of natural language for social media conditioned on the fictive authors' personality traits.*

The plan is to use the architecture and preprocessed dataset from Chapter 6 to finetune instances of both GPT-2 and ERNIE. The models will be used to generate short-text samples conditioned on the fictive author's age, gender, and personality traits. During the early phases, initial experiments on text cleaning, freezing layers, and different decoding strategies will be conducted. This is done to find the expected optimal settings and preconditions for the models' performance. Lastly, an evaluation will be conducted in line with the suggested practice for evaluating natural language generation systems (Sai et al., 2020; Van Der Lee et al., 2019) and findings from the structured literature review presented in Section 4.5. Thus combining automatic metrics and human judgment to assess the generated texts.

Note that the terms *training* and *finetuning* will be used interchangeably in this chapter. Both refer to the process of adjusting the parameters of the pretrained language models. Since the models are going to be used for natural text generation, there is no need to split the data into a training set and a test set, and hence all the data available are used for finetuning. Other terms worth describing before moving on are *generate* and *write*. Both will be used interchangeably to denote the process of the models generating text. The same applies to the terms *samples* and *generated texts*, meaning the output from the models when generating personalized natural language.

Table 7.1: Results for ERNIE 2.0 finetuned on 5 000 samples from the raw dataset with no text cleaning performed.

	Generated text
1	*!!!!!!!!!!!!!!!!!?!!!!!!!!!!!!!!!!!!!!!.!!?!!!!?!!!! ".!?i!! 1 # r * u my eyes - :?. # me me :! ya! # -! o. my a o ed ele i! # 7? : < no < > # olea # allei*
2	*time?!!!!!!!!!!!!?!!!!!!?!!!?!!!!!!!!!!!???!! "!!.?? ". "!!!? nah! “ < my me i - me " - ” -! -?! - > me my ” i, - / el mi del j me!,?. o la my y. che. j #*
3	*work thanks!!!! " ”.!! " thanks "!!!!!!!! you!!!!!..!!!!!!!!!!!!!.!!!!!? this ". thisi <,! "? " o?. - ". "? no!a -. <-?! * ” this - ni # # mr! #? <my "! <* del # # na mr!. my! my?!!*

7.2 Experimental Setup

This section is provided to give detailed information about the experimental setup in order to make the experiments repeatable and the results reproducible. As a part of the experimental setup, an initial analysis of aspects of the experiments is conducted. This analysis includes analyzing the impact of cleaning the training data, only finetuning a subset of the models' layers, the choice of training parameters, and decoding strategy for text generation. Findings from the analysis are utilized in the final experiment to achieve optimal results. It can be noted that during the setup phase, more focus is on the models' ability to generate correct grammatical text in general, examined manually. Limited attention is paid to the personalization aspect, which is left to the analysis of the final results. Before a walk-through of each part of the analysis, starting with the impact of text cleaning, the technical setup used to conduct the experiments is introduced.

7.2.1 Technical Environment

The following experiments in this project are performed on the NTNU IDUN computing cluster (Själander et al., 2019). The IDUN cluster is a part of EPIC, a computing research infrastructure at NTNU, and consists of more than 70 nodes and 90 general-purpose graphics processing units (GPGPUs). Every node in the IDUN cluster has two Intel Xeon cores with a minimum of 128 GB of main memory. The infrastructure is highly beneficial for the research because of its easy access to and use of necessary computing resources

7.2.2 Text Cleaning

The documents in the dataset to be used consist of Facebook status updates and tweets. These documents are not necessarily grammatically correct or spelled checked due to the nature of the data. A hypothesis is that some preprocessing steps should be performed before using the text as a basis to finetune the models and produce human-like text.

Table 7.2: Results from ERNIE 2.0 finetuned on 5 000 samples from the dataset with text cleaning performed.

	Generated text
1	*aires! sun blue festival the radio! - love that girls? girls! black girls coffee eyesful sweets - black - cute stupid heartsberries dusty cookies awesome doorsful -ingful tears wild blue flowers rosesing!!!aa.o school oscar beach suns? ka < " " ni ha the devil.in fr o tom beach a h no ya., beach no a da da ha h h ho angel beacheeti zaei! " the music oscari! : stars this. jorge coffee " "? a "! ojaoia] < gabriel no ya ho coffee thunder de ta mo ya the soul real you? but no a.. me. p lo ou.,e de se que ko ha no a. "*
2	*ming |! |ming mo!man.?.i jorge, a funnyness . . . o oscar tequila.o te? the team.aha!! ha me " < my birthday ni hao a.aao o thee) '!. no me mees san i yu que they de mo ou es al mo you no me loo ar ao u u you what the fuck no the life what " " oh lo ya ho " ha no es me ho vi dan? the ho o. ho - ho ta sun lo! ou ai.. - de o i ha se that la i er i!! - se oh you - no. ni / me ou es no a so? my " the house. i*
3	*|lates |? boots?ismming -doae - "o,. niti es zaja o. " de?...o...!.odoo me by meaa san thees -! -o u ye yao ya? ni i nah ha? o.hae lo se te ko e al i u es w me no " te my birthday ho eshao i?o e my coffee :e i ari te.i? p my noo santa i de the ho!!e -. lo ho. ai me. o mi ari. i ha lo me the time ta my heart!. de rama, ho y ha arie phado rt. o. oh - me i ko - - you al ja*

As a starting point for the analysis, an instance of the ERNIE 2.0 model was finetuned on 5 000 instances from a raw version of the dataset, that is, no data cleaning was performed on the text. Table 7.1 shows three examples generated from ERNIE 2.0 finetuned only on 5 000 samples when no text cleaning was performed. It can clearly be seen that none of the three examples seem human-like. The model seems to particularly have a challenge with punctuation, illustrated by the first example of Table 7.1, starting with over 50 consecutive characters with exclamation marks, question marks, and periods. Similar findings were observed for GPT-2, but not to the same extent as for ERNIE 2.0. However, the findings strengthen the hypothesis that some cleaning should be carried out.

As a consequence of the findings that both ERNIE 2.0 and GPT-2, to some degree, struggled with handling correct and human-like punctuation, cleaning the raw text is carried out. Investigation of the dataset revealed that 10% of all instances had more than triple punctuation. That is, the text (which is either a tweet or a Facebook status update) contained four or more consecutive punctuation marks. The average length of consecutive punctuation marks was 6.7 characters long. To standardize the text and avoid incorrect punctuations in the generated texts, cleaning was performed such that consecutive punctuation was limited to a maximum of three characters, as illustrated in the transformation below:

I can not believe it!!!!! The weather is terrible #rain #thunder

$$\downarrow$$

I can not believe it!!! The weather is terrible #hashtag #hashtag

Three periods or three exclamation marks placed at the end of a sentence are more likely to be a conscious choice and are thus kept in the texts. As seen in the example above, all hashtags are replaced by a generic *#hashtag* string and all hyperlinks by *URL* during the cleaning process. Recall from Chapter 5 that the same was already present for user mentions in the existing datasets. The reason for the replacement of hashtags and hyperlinks as well is to standardize the text even more. Table 7.2 shows generated samples from ERNIE 2.0 finetuned on the cleaned text and reveals a significant improvement from the samples in Table 7.1.

7.2.3 The Effect of Freezing of Layers

Finetuning transformer-based models can be an intense and resource-consuming task. An architectural choice was made in Chapter 6 to freeze 25% of the models' layers and finetune only the other layers. The reason for this choice of freezing layers is that both ERNIE and GPT-2 are already pretrained on data available on online websites. A hypothesis is that only finetuning a subset of the layers can speed up the finetuning process without losing remarkable quality in the results.

To test the hypothesis that layers can be frozen without losing remarkable quality and verify the architectural choice, two instances of both ERNIE and GPT-2 are finetuned on a subset of the dataset. For one of the instances of each model, the last 25% of the layers are frozen, which means that parameters in those layers will not be updated during the training phase. Since this is a part of the initial experiments for optimizing the setup, a random subset of the processed dataset containing 3 000 samples is used for finetuning. The same subset is used for both the models so that it is possible to compare the results. All parameters for the training are set to the default values.

When manually examining the preliminary results, no observable differences in the produced samples were revealed. Hence the choice of keeping a subset of the layers frozen was considered appropriate due to the benefits during training and generation.

7.2.4 Parameters for Finetuning

The parameters used for finetuning both GPT-2 and ERNIE 2.0 are placed in Table 7.3. The related work from Chapter 4 that used either GPT or GPT-2, Zhang et al. (2020) and Golovanov et al. (2019), did not report their parameter settings. Regards the learning rate, GPT (Radford et al., 2018), which GPT-2 extends, used a learning rate of $5e^{-5}$ for their finetuning. Sun et al. (2020) also reported using a learning rate of $5e^{-5}$. Hence a learning rate of $5e^{-5}$ was considered appropriate. The same applies to the choice of an

Table 7.3: The parameters used for finetuning the models.

Parameter	**Value**
Batch size	8
Learning rate	$5e^{-5}$
Optimizer	Adam
Adam β_1	0.9
Adam β_2	0.999
Adam ϵ	$1e^{-8}$
Number of training epochs	3.0

Table 7.4: The parameters used for text generation.

Parameter	**Value**
Top K	50
Top p	0.95
Max length	300

optimizer, where the Adam optimizer was used by both Radford et al. (2018) and Sun et al. (2020). The other parameters specifying the optimizer are solely the default values provided. Radford et al. (2019) reports three epochs to be sufficient when finetuning. Sun et al. (2020) used three and four epochs for the experimental finetuning tasks. Thus three epochs were also considered a good choice for the finetuning in this thesis.

7.2.5 Decoding Strategy for Generation

In text generation, the decoding strategy describes how the model chooses the next token to write. Using a greedy strategy, the model would always choose the next word with the highest probability. In this way, the model can skip word sequences with higher probabilities and be stuck in repeating loops since it never looks more than one word ahead. To avoid repetitive generation, a combination of Top-K sampling (Fan et al., 2018) and Top-p sampling is used at the decoding strategy, and Table 7.4 shows the parameter settings used.

The value for K in Top-K sampling says that when the model will choose the next word in, the K most probable words are extracted, and the next token would be chosen among those K words. With Top-p, on the other hand, a probability p is specified. A minimum set of tokens with their probability sum up to p is extracted, and the next token is then chosen from this set. Top-K and Top-p in combination can help to avoid words with very low probability but at the same time allowing some flexibility. The maximum length is set to 300 characters based upon the maximum length of a Tweet is 280 characters, and that Facebook allows for longer, 300 characters were considered a proper maximum length.

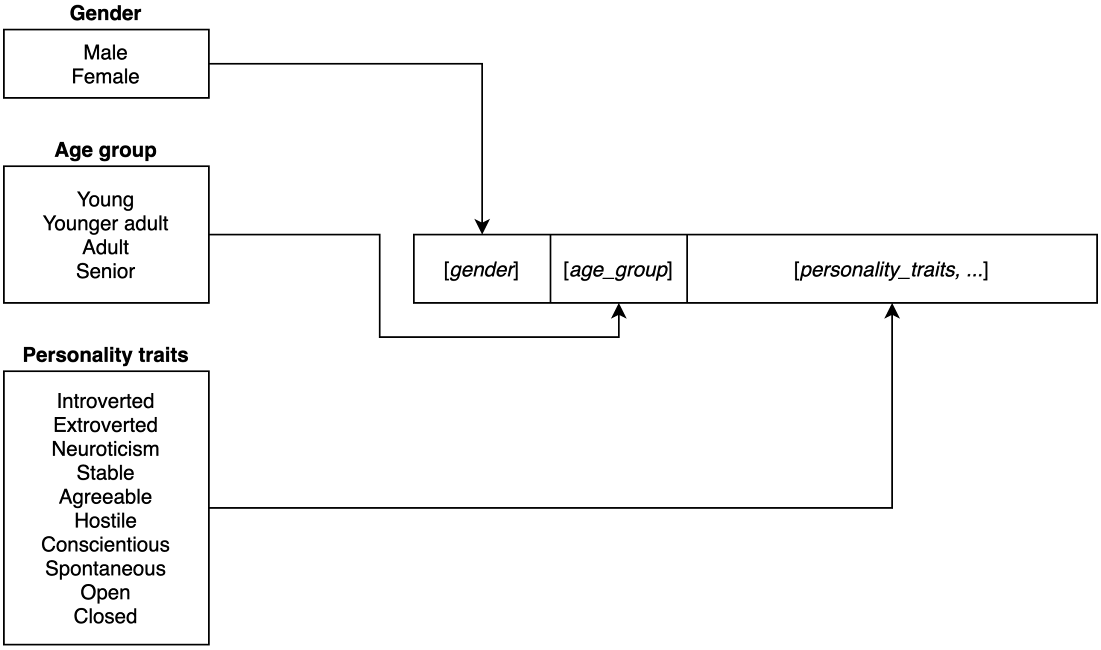


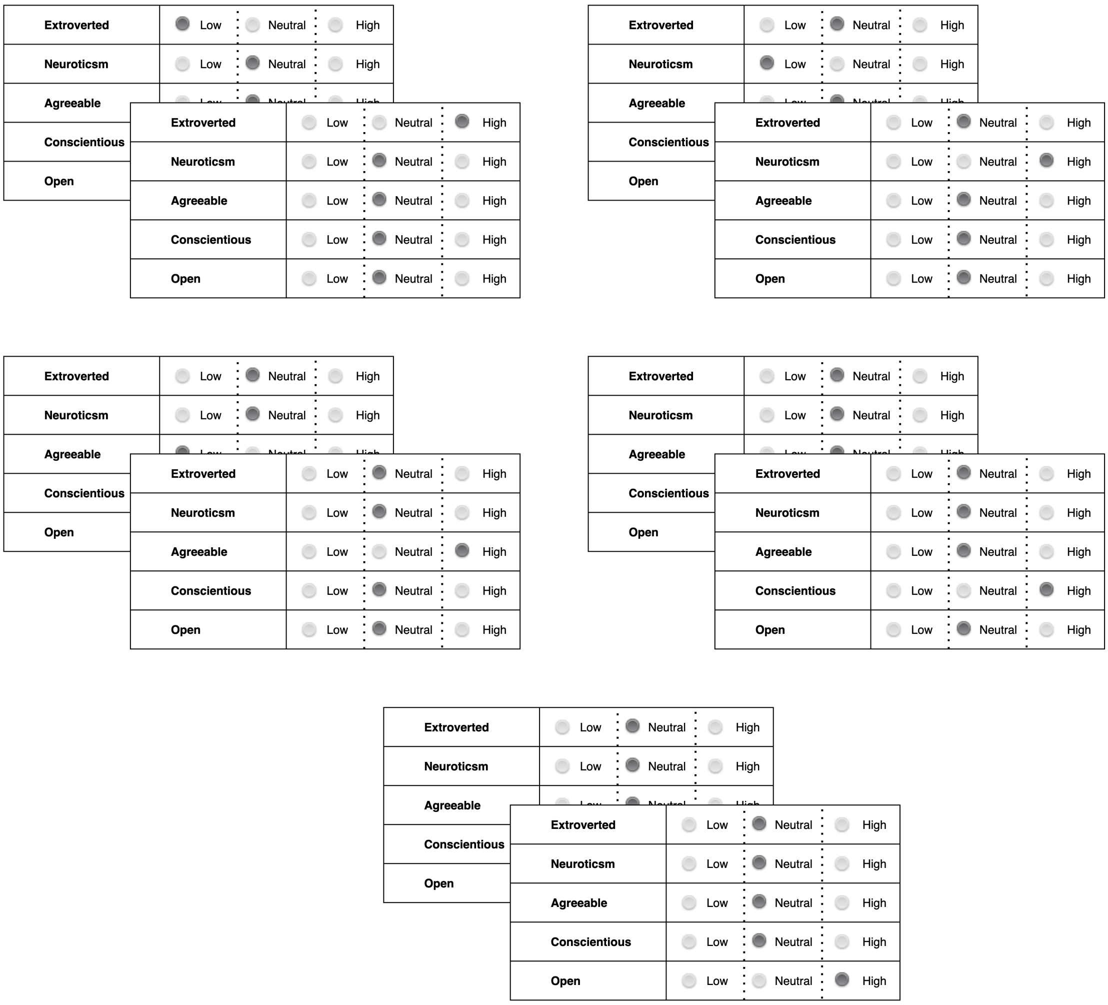
Figure 7.1: The different conditional input settings for generation using the keyword version.

7.2.6 Conditional Input Settings

The architecture from Chapter 6 is used to generate text from the finetuned models, which is the fundamental basis for the results. This section describes the conditional input parameters used to generate the texts for in the final results. Recall that the architecture is designed to support two different specifications of conditional inputs. The keyword format, where a number of keywords describing the fictive author are specified, and the personality trait version, where the degree of all the five personality traits of the Big Five model are set. Both versions support personalization with respect to personality traits, but only the first one combines it with age and gender as well. These different conditional input settings are describing attributes of fictive authors, which the models should mimic when generating samples.

The keyword format allows for numerous possibilities, illustrated by Figure 7.1. For this experiment, it is chosen to generate samples for both genders in combination with all age groups, but only one personality trait keyword. For example: *female*, *young*, and *open*. The architecture supports input of multiple personality trait keywords (extending the example to: *female*, *young*, *open*, *introverted*, *stable*). However, it is considered most feasible and manageable during the evaluation to specify only one keyword describing personality. Nonetheless, note that it would not make sense to input two opposite traits, for instance *introverted* and *extroverted*, since these are in direct contrast to each other.

The personality trait solution also allows for numerous input settings, Having five



Extroverted	● Low	○ Neutral	○ High
Neuroticsm	○ Low	● Neutral	○ High
Agreeable	○ Low	● Neutral	○ High
Conscientious			
Open			

Extroverted	○ Low	○ Neutral	● High
Neuroticsm	○ Low	● Neutral	○ High
Agreeable	○ Low	● Neutral	○ High
Conscientious	○ Low	● Neutral	○ High
Open	○ Low	● Neutral	○ High

Extroverted	○ Low	● Neutral	○ High
Neuroticsm	● Low	○ Neutral	○ High
Agreeable	○ Low	● Neutral	○ High
Conscientious			
Open			

Extroverted	○ Low	● Neutral	○ High
Neuroticsm	○ Low	○ Neutral	● High
Agreeable	○ Low	● Neutral	○ High
Conscientious	○ Low	● Neutral	○ High
Open	○ Low	● Neutral	○ High

Extroverted	○ Low	● Neutral	○ High
Neuroticsm	○ Low	● Neutral	○ High
Agreeable	● Low	○ Neutral	○ High
Conscientious			
Open			

Extroverted	○ Low	● Neutral	○ High
Neuroticsm	○ Low	● Neutral	○ High
Agreeable	○ Low	○ Neutral	● High
Conscientious	○ Low	● Neutral	○ High
Open	○ Low	● Neutral	○ High

Extroverted	○ Low	● Neutral	○ High
Neuroticsm	○ Low	● Neutral	○ High
Agreeable	○ Low	● Neutral	○ High
Conscientious			
Open			

Extroverted	○ Low	● Neutral	○ High
Neuroticsm	○ Low	● Neutral	○ High
Agreeable	○ Low	● Neutral	○ High
Conscientious	○ Low	○ Neutral	● High
Open	○ Low	● Neutral	○ High

Extroverted	○ Low	● Neutral	○ High
Neuroticsm	○ Low	● Neutral	○ High
Agreeable	○ Low	● Neutral	○ High
Conscientious			
Open			

Extroverted	○ Low	● Neutral	○ High
Neuroticsm	○ Low	● Neutral	○ High
Agreeable	○ Low	● Neutral	○ High
Conscientious	○ Low	● Neutral	○ High
Open	○ Low	○ Neutral	● High

Figure 7.2: The different conditional input settings for generation using the the personality traits version.

personality trait settings with three options for each (*low*, *neutral*, and *high*), giving over 200 combinations. It was decided that the best procedure for this experiment was to for each personality trait, generate samples for setting the trait to both *low* and *high* while keeping the other traits on the *neutral* setting. This gives ten different input settings, illustrated by Figure 7.2.

7.3 Experimental Results

The experimental setup is used to generate a number of samples per model and input format version. For GPT-2, the number of samples generated per input combination is 100. For ERNIE 2.0, on the other hand, the number of samples is limited to 20 per input setting. The reason for this is a manual observation during the development that ERNIE 2.0 consistently generated much longer and incomprehensible samples. Even though it is generated fewer samples per setting for ERNIE 2.0, the total data produced per model does not differ much since GPT-2 produced much shorter texts. All the samples are analyzed, and the results will be presented in this section. A subset of the generated samples from both GPT-2 and ERNIE 2.0 can also be found in Appendix D.

According to the experimental plan from Section 7.1, statistical analyses of the generated samples are carried out. Knowledge about expected writing styles for certain characteristics presented in Chapter 3 is used to analyze whether the generated samples align with expectations. The results will be evaluated and discussed in Chapter 8.

7.3.1 Results from Human Evaluation

Human evaluation of the generated samples was conducted as a part of the evaluation procedure. The total number of human judges who participated by answering a questionnaire was 26. The questionnaire (showed in Appendix E) was developed and all the judges assessed the same generated samples. First, 24 generated texts were evaluated by the judges on the fluency of each text and whether each text made sense. Secondly, for ten other generated samples, the judges were asked to guess the attributes (age group, gender, and personality trait) of the author of each of the samples.

Fluency and making sense was measured on a scale from one to five, where five indicates the best value for both metrics. The mean and standard deviation of both fluency and making sense for all assessed samples are placed in Table 7.5. In Table 7.6, the scores are averaged per model. In both, the versions of GPT-2 and ERNIE 2.0 are treated as a whole and are not separated per input version. This is considered the right choice in order to limit the number of samples for the judges to assess and to avoid a too extensive questionnaire some judges might not answer fully. In accordance, it is presumed that the finetuned models' ability to write a fluent text that makes sense is not dependent on the conditional input format.

As mentioned above, the second part of the human evaluation was the judges guessing the age group, the gender, and the personality trait of each sample's author. The evaluation form did not specify whether the samples were written by a human or not.

Table 7.5: The mean score and standard deviation on fluency and making sense. Samples with IDs 8, 15, 21, and 24 are generated by ERNIE 2.0. The other samples are generated by GPT-2.

ID	**Fluency**		**Making sense**	
	Mean	**SD**	**Mean**	**SD**
1	2.69	1.38	1.81	0.80
2	4.85	0.37	4.58	0.95
3	3.15	1.46	2.65	1.52
4	2.73	1.28	1.85	0.88
5	3.65	1.06	2.46	1.17
6	3.58	1.03	2.96	1.11
7	3.31	1.09	2.50	0.91
8	1.12	0.33	1.04	0.20
9	3.04	0.92	2.46	0.95
10	4.85	0.61	4.69	0.79
11	4.31	1.01	4.19	0.98
12	3.58	0.90	3.42	1.03
13	4.69	0.55	4.58	0.76
14	4.62	0.70	4.62	0.94
15	1.15	0.37	1.04	0.20
16	3.12	1.14	3.04	1.04
17	3.69	0.88	2.65	1.13
18	4.96	0.20	4.92	0.27
19	3.77	0.95	2.81	0.94
20	4.96	0.20	4.92	0.27
21	1.15	0.37	1.12	0.33
22	4.85	0.37	3.96	1.11
23	4.19	0.57	3.85	0.97
24	1.15	0.37	1.04	0.20
Average	3.46	0.75	3.05	0.81

Table 7.6: The average mean and standard deviation per model on the human evaluation of fluency and making sense.

	Fluency		**Making sense**	
	Average mean	**Average SD**	**Average mean**	**Average SD**
GPT-2	3.93	0.83	3.45	0.93
ERNIE 2.0	1.14	0.36	1.06	0.23

Confusion matrices for the human predictions regards age group and gender can be found in Figure 7.3. Note that due to the observation of what seemed like a low quality of the ERNIE 2.0 texts, this human prediction of author attributes was only performed for samples generated by GPT-2.

The true labels on the y-axis of Figure 7.3 mean that the text was generated with those values as conditional inputs. Since language models have generated all the samples, there is no actual human with real attributes who wrote the texts. So the true values in this manner are the actual conditional input value. If all predictions were made correctly, the diagonal from upper left to lower right would be colored dark blue since there would be a full match between the judges' guesses and the actual conditional inputs.

Figure 7.3a regarding age group shows that it was the best match between the conditional input and judges' guess for the young age group. The judges guessed the senior group in general seldom. It is noteworthy that text written by all the age groups except for the youngest was most often characterized as written by an adult. Regards gender, illustrated in Figure 7.3b, females were most often guessed correctly. On the other hand, the males' text was predicted correctly in 54% of the cases, which makes almost the same as a random guess.

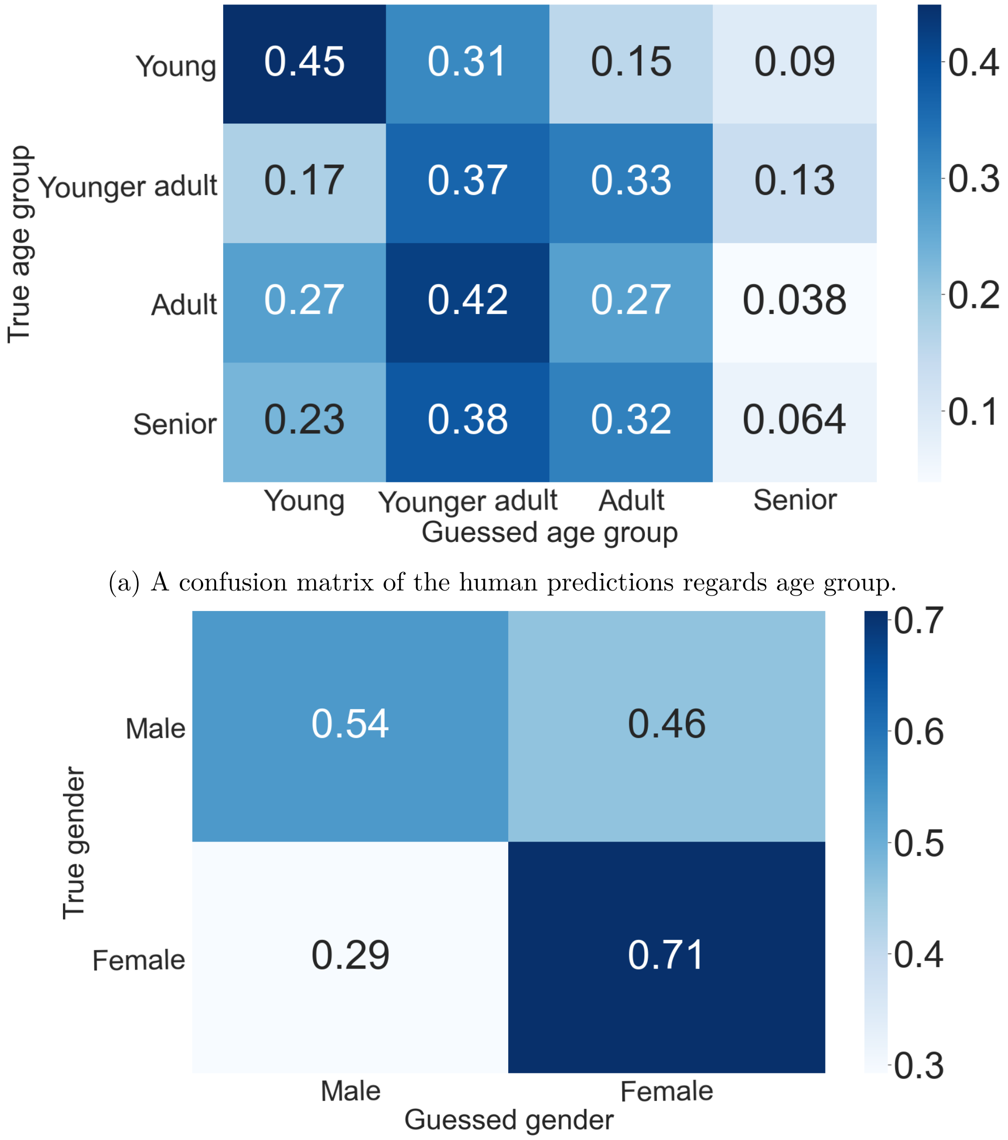
The results of the human predictions concerning personality can be found in Figure 7.4a, 7.4b, Figure 7.5a, 7.5b, and Figure 7.6. The results are separated per pair of personality traits. Regards the *neither* choice in the figures, choosing neither was an alternative for all samples when predicting the personality traits, which can be seen in the evaluation form in Appendix E. Of all the five pair of personality traits, the stable and neuroticism trait in Figure 7.6 was most predicted correctly. For extroverted and introverted seen in Figure 7.5a, the human predictions were skewed to the extroverted trait. The agreeable samples were clearly more predicted as hostile texts, shown in Figure 7.5b. For both pairs of open and closed (Figure 7.4a), and spontaneous and conscientious (Figure 7.4b), the predictions are shifted to favoring the former alternative.

7.3.2 Experimental Results Concerning Personality

The next part of the evaluation will present the results of to what degree personality traits are preserved and present in the generated texts, and show the potential differences between GPT-2 and ERNIE 2.0. Keep in mind that the specification of personality traits is supported by both the keyword input and personality traits input version. First, the results from both GPT-2 and ERNIE using the keyword input format are presented, then the results from the personality trait input follow. Throughout this section, the phrase *the introverted text* is meant to target the samples generated with the introverted trait as a conditional input and similar for the other personality traits.

Keyword Input Format

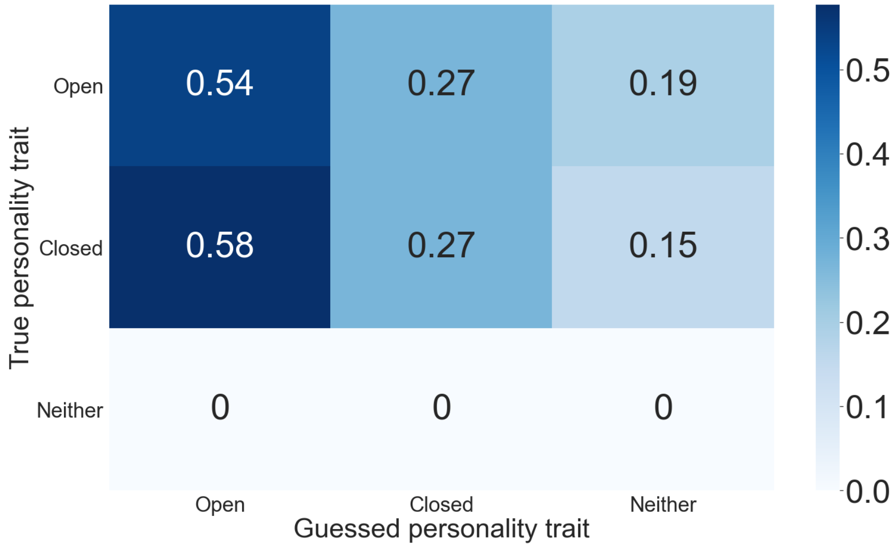
The generated samples from both the GPT-2 model and ERNIE 2.0 model are grouped by personality traits. Recall from Section 7.2.6 that it was chosen to generate samples with only one keyword specifying personality, hence the produced samples can be grouped



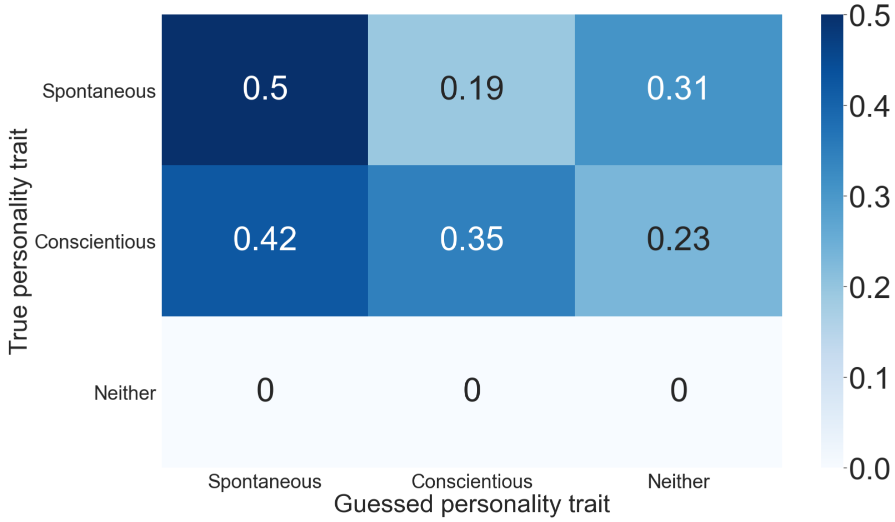
(a) A confusion matrix of the human predictions regards age group.

(b) A confusion matrix of human predictions regards gender.

Figure 7.3: The results of the human predictions for age group and gender.

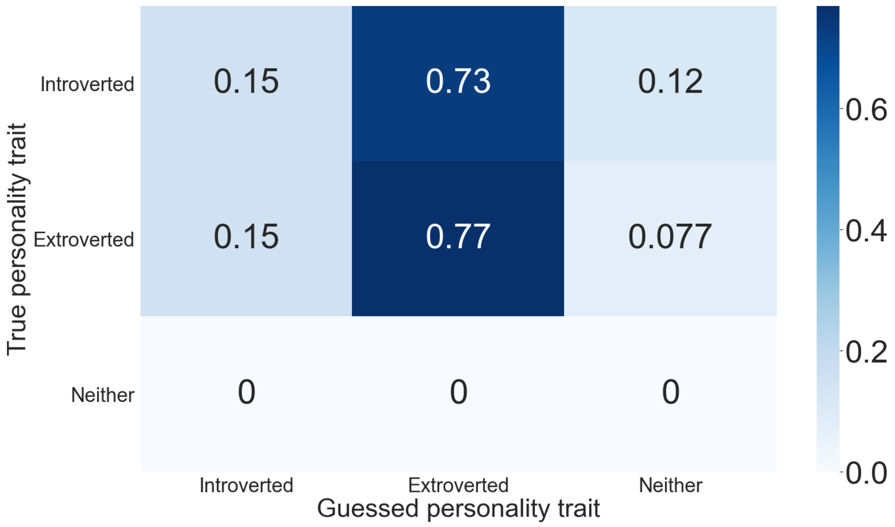


(a) A confusion matrix of the human predictions for the open and closed personality traits.

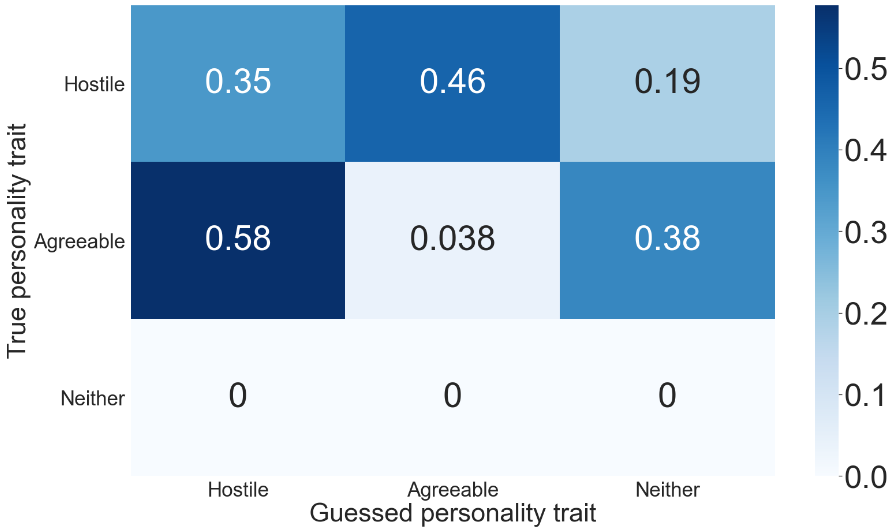


(b) A confusion matrix of the human predictions for the conscientiousness and spontaneous personality traits.

Figure 7.4: The results of the human predictions for the personality traits open and closed, and conscientiousness and spontaneous.



(a) A confusion matrix of the human predictions for the introverted and extroverted personality traits.



(b) A confusion matrix of the human predictions for the agreeable and hostile personality traits.

Figure 7.5: The results of the human predictions for the personality traits introverted and extroverted, and agreeable and hostile.

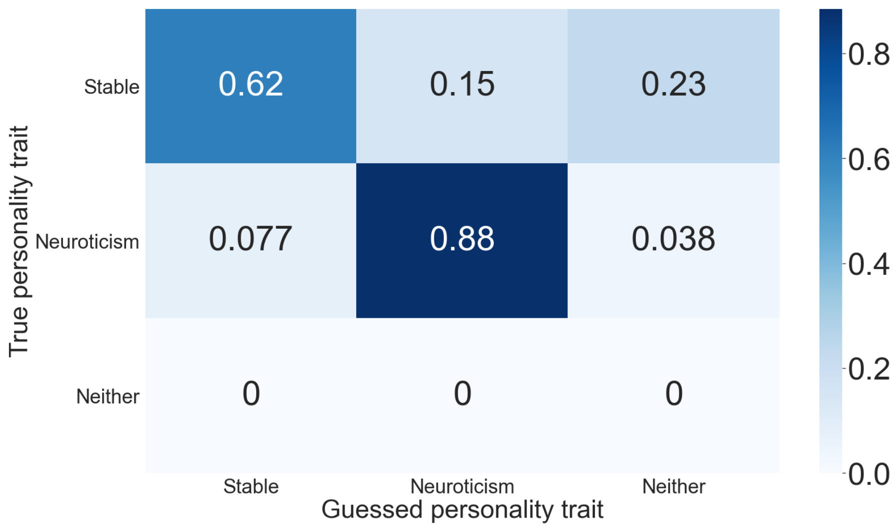


Figure 7.6: A confusion matrix of the human predictions for the neuroticism and stable personality traits.

according to the specified personality trait. Figure 7.7 and Figure 7.8 show a heatmap of the normalized means of 12 features' occurrences in the results from GPT-2 and ERNIE 2.0 respectively.

These features are chosen according to findings from Section 3.5 about characteristics of written text that can represent differences between personality traits. The normalized mean value of the features is chosen because it measures how frequently each feature occurs per personality trait relative to the other traits. The mean is taken to make the values comparable between GPT-2 and ERNIE 2.0. The values are then normalized due to easy comparison of occurrence across the different features. The y-axis in Figure 7.7 and Figure 7.8 shows the personality traits, and the different features computed are distributed among the x-axis, forming a grid. The darker color in a square, the greater the value for that feature in the samples with the keyword.

When examining the heatmap for GPT-2 in Figure 7.7, it can be seen that texts generated with neuroticism stand out with the shortest texts, illustrated by a low score for the number of words, number of characters, and the total length. When it comes to punctuation, the open texts have the most number of exclamation marks, and extroverted texts the most number of question marks. In total, the agreeable samples also have most occurrences of punctuation in general, followed by the extroverted texts. The traits conscientious and agreeable have the most articles (*a*, *an*, and *the*). Regards the use of swear words, stable samples have the fewest and closed samples the most. According to GPT-2, the word *you* is most used in open texts, followed by text generated with agreeable as the conditional input.

The results produced by the ERNIE 2.0 model are illustrated in the heatmap in Figure 7.8. The closed samples are the shortest and have the fewest words, this is seen by the closed trait having the lowest value for the number of words, the number of

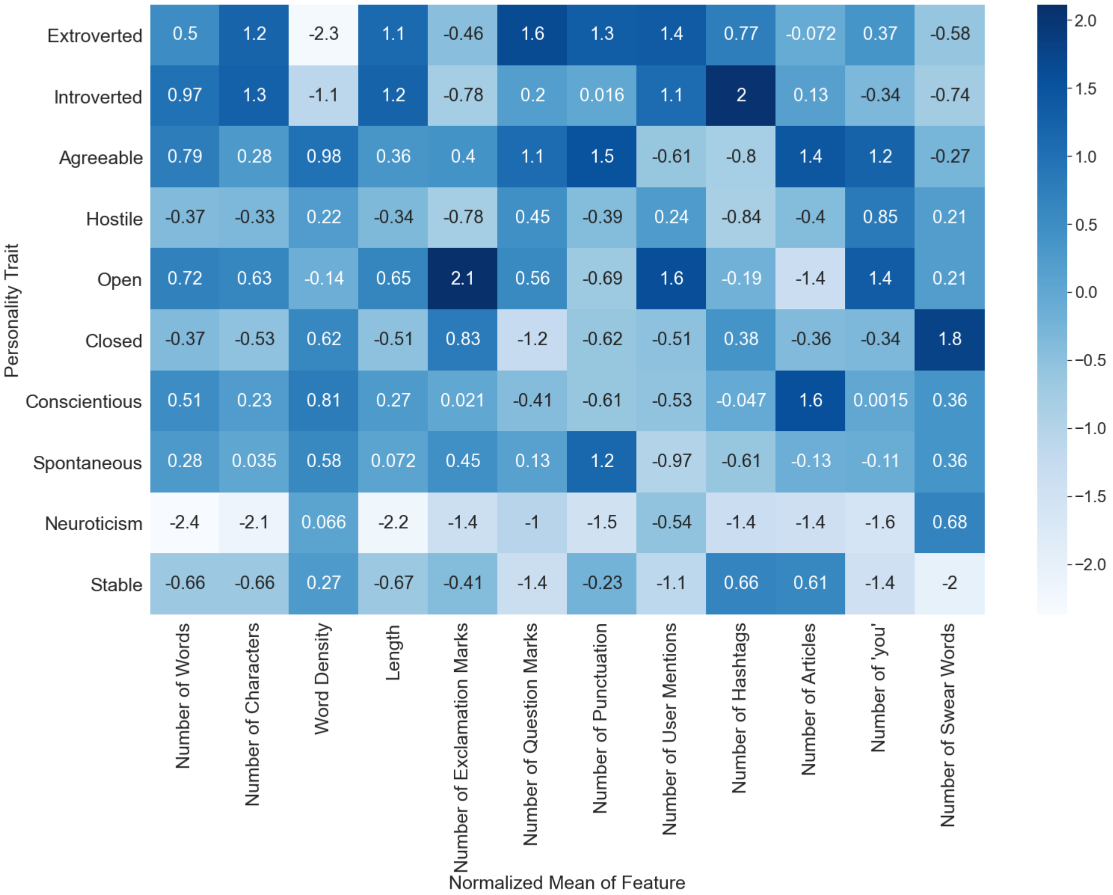


Figure 7.7: A heatmap of normalized means of occurrences per generated sample of features with respect to personality traits for the GPT-2 model keyword version.

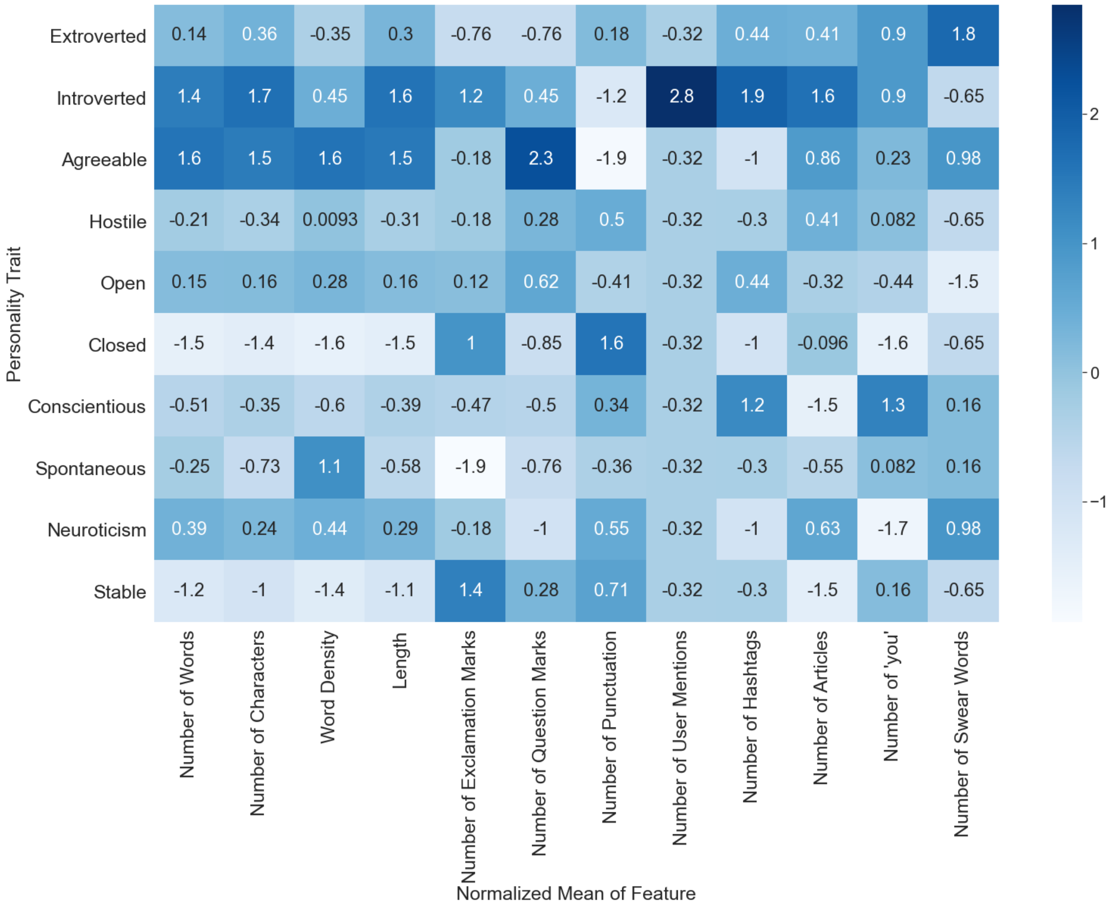


Figure 7.8: A heatmap of normalized means of occurrences per generated sample of features with respect to personality traits for the ERNIE 2.0 model keyword version.

characters, and length. The stable samples, followed by the introverted texts, have the most occurrences of exclamation marks. Whereas the question mark is most used in the agreeable samples. From the ERNIE 2.0 model, the introverted texts are the only ones containing user mentions, hence the equal value for all the other traits. Hashtags and articles are also most present in the introverted samples. The word *you* is least used by samples with neuroticism as the input keyword. Regards the occurrences of swear words, this is most present in extroverted texts.

Personality Trait Input Format

Similarly to the keyword input format, heatmaps for both GPT-2 and ERNIE 2.0 on the personality input trait format are shown in Figure 7.7 and Figure 7.8. The personality trait input format supported specifying *low*, *neutral*, or *high* on each of the Big Five personality traits, this was explained in Section 6.2.5. Even though the input strictly speaking specify low, neutral, or high on the traits, scoring high on a trait is the same as scoring low on the opposite personality trait. To illustrate, extroverted and introverted are opposites. According to the Big 5 model, scoring low on extroverted corresponds to being introverted. Since it is more convenient to present and discuss the results using the opposite terms (*introverted*) for the traits rather than the low score (*low on the extroverted scale*), the results are presented using this terminology.

Figure 7.7 illustrates the results from the GPT-2 model using personality trait input. The number of words and characters and the total length of the samples are highest for the hostile samples. Text generated with the trait stable stands out at the one with the lowest word density, that is the number of words divided by the number of characters. The higher the word density value, the larger the number of words per character, implying more and shorter words. Stable also stands out with the most use of exclamation marks, whereas the open texts have most question marks. An interesting notice is that samples generated mimicking a hostile person have the highest number of swear words.

The results for ERNIE 2.0 personality trait input version are shown in the heatmap in Figure 7.8. The stable texts have the most number of words, whereas the introverted ones have the most characters. The introverted samples also have the longest texts and the highest numbers of exclamation marks. Question marks appear the most in the open samples. Regards the number of punctuation, the extroverted texts stand out with the least use of punctuation. User mentions only appear in the neuroticism samples, hence all the other traits have the same value with respect to the number of user mentions. Articles are least used in the closed samples, and the word *you* is most used by text generated with the neuroticism trait. Swear words occur in general seldom, and very few occurrences in some traits can affect the normalized means to make it look like there is a more significant difference. This fact is not directly visible in the heat map. However, this why extroverted, hostile, and neuroticism have the same value with respect to the number of swear words and corresponding for the rest of the traits.

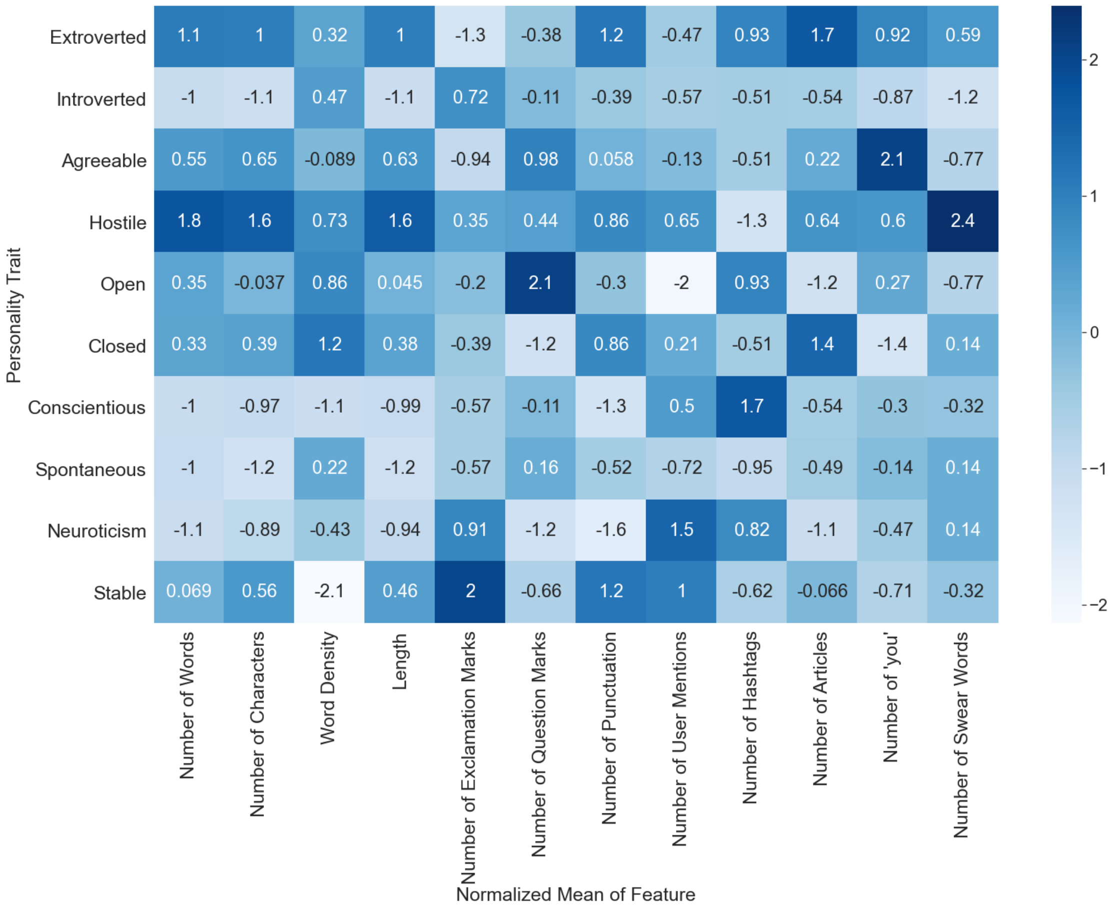


Figure 7.9: A heatmap of normalized means of occurrences per generated sample of features with respect to personality traits for the GPT-2 model personality trait version.

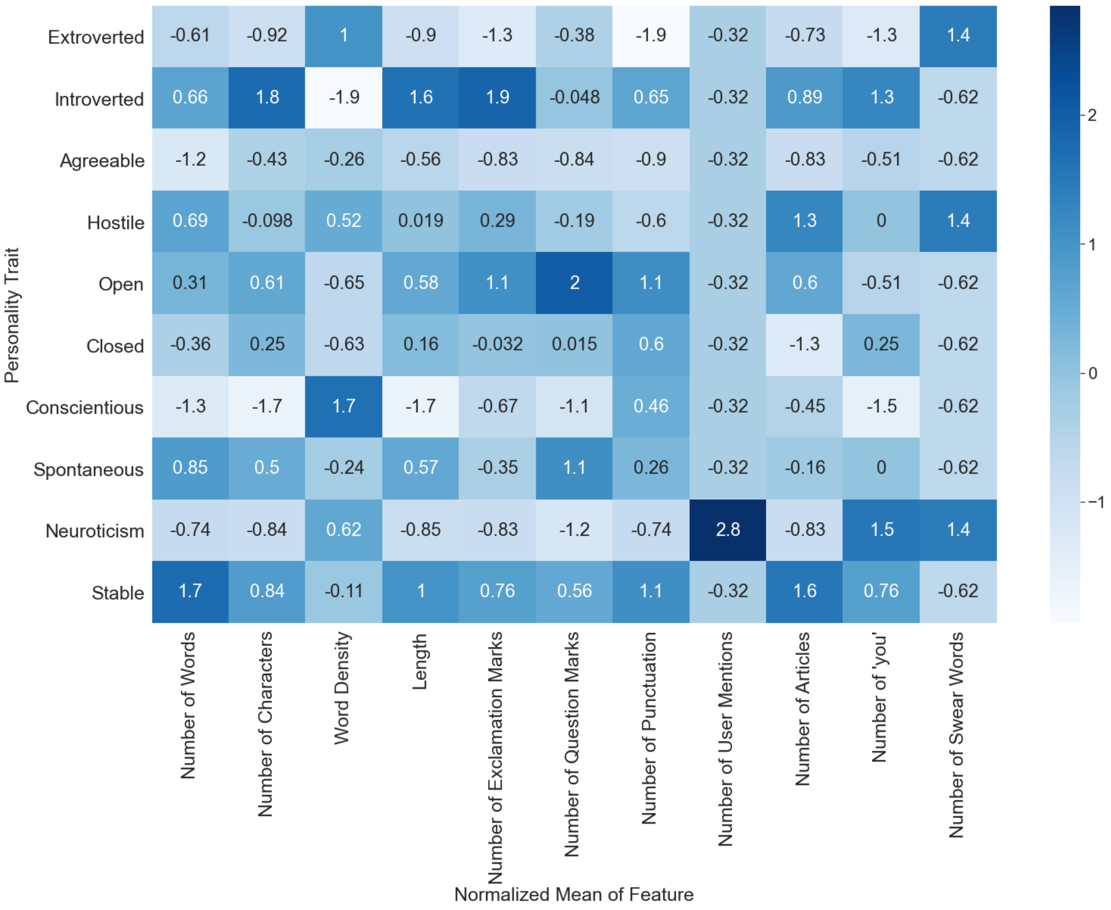


Figure 7.10: A heatmap of normalized means of occurrences per generated sample of features with respect to personality traits for the ERNIE 2.0 model personality trait version.

Table 7.7: The mean of occurrences per generated sample of features with respect to the gender attribute

	GPT-2		ERNIE 2.0	
Feature	**Male**	**Female**	**Male**	**Female**
Articles	0.97	0.92	1.25	1.18
You	0.64	0.65	2.09	1.96
Negations	0.29	0.28	1.22	1.30
Pronouns	2.35	2.36	11.31	11.41
Swear words	0.03	0.03	0.03	0.02

7.3.3 Experimental Results Concerning Gender

The results with respect to gender are conducted the same way as for personality by analyzing occurrences of relevant features. Recall that controlling the text generation by the gender attribute is only available using the keyword format of conditional input, hence the trait input version is irrelevant in this section. Since this section handles only two genders (male and female) versus ten different personality traits, an extensive description of the results is more suitable than illustrating with a heatmap.

This section will present the results of both GPT-2 and ERNIE 2.0 when analyzing findings with respect to the gender attribute. All generated samples for each model are grouped by gender, and analyses of the two groups are carried out. Note that the terms *produced samples*, *generated texts*, and *written text* will be used interchangeably to talk about the text the models generate during the conditional generation. In the same way, female texts and male texts will be used to address text produced with respectively female and male as input keywords.

The produced samples from GPT-2 with respect to gender differ minimally on the number of words, characters, and the total length of the texts. According to GPT-2, the difference from female to male is less than 1% for those features between the female and the male texts and is thus interpreted as negligible. The same applies to the produced samples by ERNIE 2.0 for the same features, with less than 1% difference from female to male.

However, looking at the number of question marks and exclamation marks in the generated samples, GPT-2 and ERNIE 2.0 disagrees. GPT-2 wrote samples with females having a 21.8% increase in the number of question marks compared to males. ERNIE 2.0, on the other hand, produced male samples with an increase of 13.1% number of exclamation marks compared to the female texts.

Regards the number of articles, occurrences of the word *you*, number of negations (words like *no*, *don't*, and *isn't*), number of pronouns (*I*, *you*, *him* and following words), and the use of swear words, Table 7.7 shows the mean for each of the model computed per gender.

When it comes to the presence of user mentions, hashtags, and hyperlinks in the generated texts, GPT-2 produced samples with minor differences between the genders,

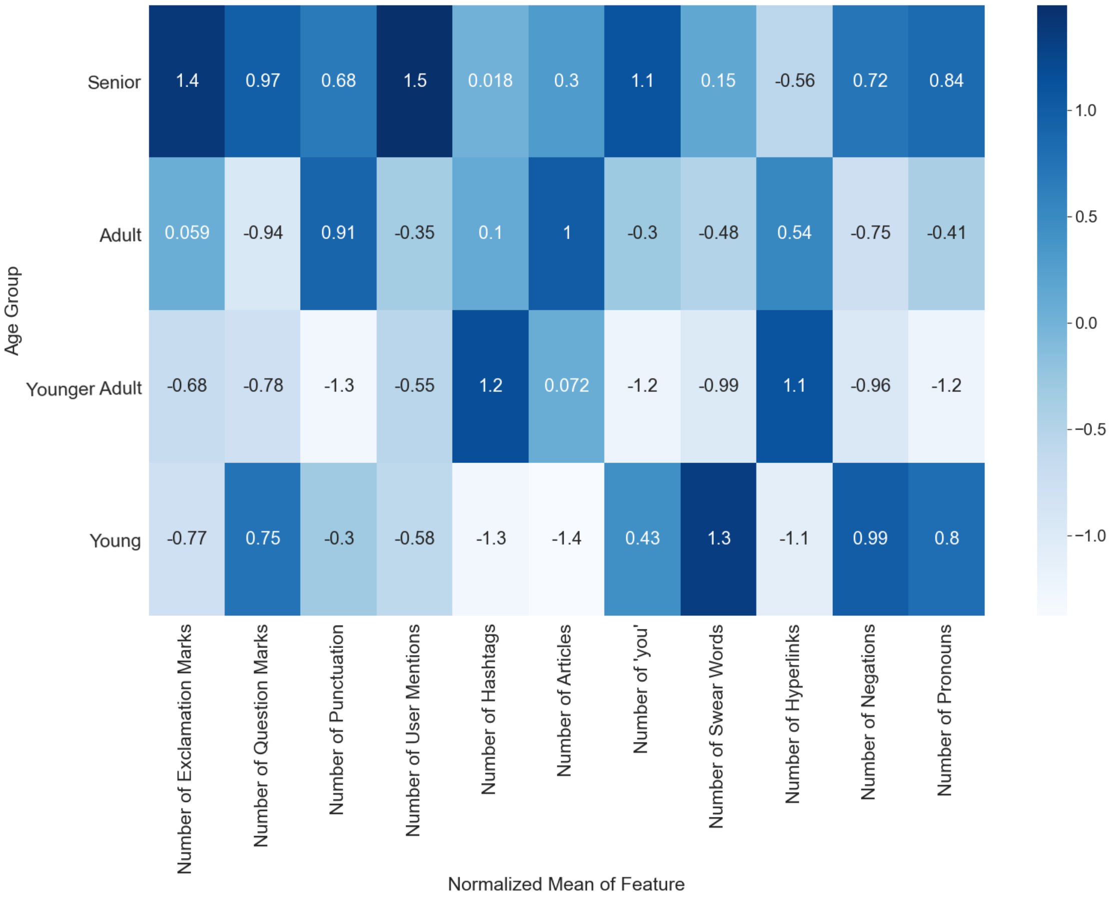


Figure 7.11: A heatmap of normalized means of occurrences per generated sample of features with respect to the gender attribute for GPT-2.

and ERNIE 2.0, in general, used a few of these features. For GPT-2, female texts had 3.1% more user mentions than male texts. The count of hashtags was 2.1% more in the male texts, and males used 3.7% more hyperlinks. The samples written by ERNIE 2.0 contained so few (less than ten occurrences) of each of these features, being considered too few to conclude between the male and female samples.

Regards emoticons, the heart emoticon (*<3*) was present 128.1% more in the female texts. Emoticons with a hyphen, interpreted as a nose (*:-)*), occurred 116.67% more in the female samples. Emoticons without a nose (*:)*) differed less between the genders and was only 4.8% more in the male texts. ERNIE 2.0 produces no samples containing emoticons, thus no difference between the genders can be observed.

7.3.4 Experimental Results Concerning Age Group

Similarly as for personality and gender, the samples are also grouped per age group and analyzed accordingly. As for gender, the conditional input for the age group is only

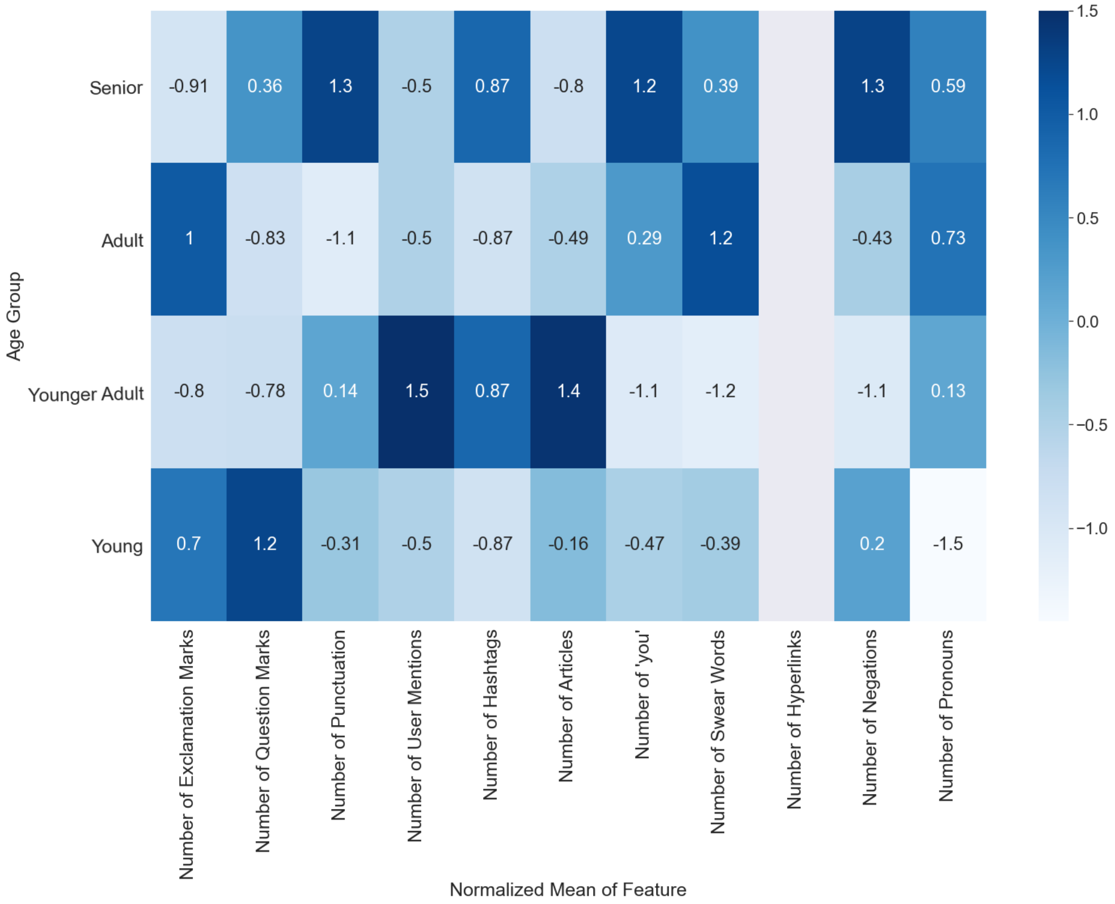


Figure 7.12: A heatmap of normalized means of occurrences per generated sample of features with respect to the gender attribute for ERNIE 2.0.

available for the keyword input version. Heatmaps for the occurrence of features for GPT-2 and ERNIE 2.0 are shown in Figure 7.11 and Figure 7.12, respectively.

For GPT-2, the seniors clearly use the most exclamation marks and question marks. However, punctuation in general appears the most in the adult age group. User mentions are least used by the youngest, and the use increase with older age. Swear words also occur the most in the youngest texts. Text generated with younger adult as the age group uses the most hashtags and hyperlinks. Both hashtags and pronouns are most used in the young and senior groups and less in the younger adult and adult groups.

When looking at the results from ERNIE 2.0 in Figure 7.12, it can first be noted from the empty column that no hyperlinks appear in the generated samples. Otherwise, exclamation marks occur most in the adult age group, whereas question marks are most used in the young age group. However, punctuation in general occurs the most in the senior texts. Pronouns are less used in the young texts, and negations occur the most in the senior samples. Overall, there are few hashtags in the samples, which is the reason for the equal values between the young and adult groups, and the younger adult and the senior groups.

8 Evaluation and Discussion

The first part of this chapter will evaluate the results presented in Chapter 7. It will concern the quality of the generated texts concerning fluency and coherence, and whether personality traits, genders, and age groups are reflected in the generated samples. The second part of the chapter will discuss the results and findings in the light of their implications, unite the existing literature with the obtained results, and clarify limitations in the research conducted.

8.1 Evaluation

This section evaluates the results presented in Section 7.3. First, considering the fluency and whether the generated texts make sense, which is assessed by human judges. Then, an evaluation with respect to the three aspects of human attributes: personality, gender, and age are carried out. Do the generated samples reveal characteristics that are found in the literature to correspond with specific author attributes?

8.1.1 Evaluation of Results Concerning Fluency and Making Sense

The result from the human assessment of the generated samples shows a significant difference between GPT-2 and ERNIE 2.0 on both fluency and making sense. Table 7.5 shows the mean and standard deviation for each rated sample on both criteria. In Table 7.6, the scores are distributed per model and averaged, giving a summarized view of the overall score per model. From this, it can clearly be seen that there is a significant difference between GPT-2 and ERNIE 2.0, and that GPT-2 are rated notably better on both fluency and making sense.

According to the average of the humans who participated, the results from GPT-2 are promising, with good fluency and making somewhat sense. On the other hand, the results from ERNIE 2.0 are unsatisfactory, scoring very poorly on grammatical correctness and making no sense at all, according to the judges. It can also be seen that the samples from ERNIE 2.0 obtained a lower standard deviation for both fluency and making sense than GPT-2. Hence, the human judges agreed more on the low scores for ERNIE 2.0.

Since GPT-2 shows the most promising results, a further qualitative look at those samples is taken. Noticeably, the samples with ID 18 and 20 obtained the highest mean score on fluency and making sense. These samples are corresponding:

> *I feel bad for you.*

and:

> *I'm sorry.*

The short length of these generated samples could be a reason for the high scores, shown by the lowest standard deviations, which indicates a high agreement among the judges. Naturally, the shorter text, the fewer grammatical errors are possible to make and hence somewhat easier to write grammatically correct. Looking on the other side, the sample from GPT-2 with the lowest score on both fluency and making sense is the following with ID 1:

> *MySpace, but I'm serious about having a conversation with you would be if your girlfriend has something to say about it.*

The sample is generated with correct grammatical words, and the sentence structure is partly sufficient concerning the ordering of the words. However, the message cannot be directly understood and makes no immediate sense.

To summarize the evaluation of the fluency and whether the texts make sense, there is a significant difference between the models. GPT-2 is shown to be promising in writing grammatically correct texts that also make sense. For ERNIE 2.0, on the other hand, the human judges agreed that those samples were not satisfactory on either fluency or making sense.

8.1.2 Evaluation of Results Concerning Personality

The next part of the evaluation concerns whether the generated samples are shown to reflect personality traits. This is not a straightforward task but recall from Section 3.5 that correlations between social media text and the authors' personality traits are evident. These findings can be used to evaluate whether the generated texts align with expectations regards personality traits by examining the experimental results from Section 7.3.2. Features describing characteristics are extracted from the generated texts, and the normalized mean of the occurrence of each feature per each personality trait is computed. The normalized mean value is chosen because it measures how frequently each feature occurs per personality trait relative to the other traits. The mean is taken to make the values comparable between GPT-2 and ERNIE 2.0. The values are then normalized due to easy comparison of occurrence across the different features.

The following characteristics from Section 3.5 cover all five dimensions of the Big Five model, and the generated texts will be analyzed accordingly. The extroverted trait is shown to correlate with more words per tweet. Persons scoring high on agreeable use the word *you* more often. Swear words seems more used by spontaneous users. The use of exclamation marks correlates with the trait neuroticism, and the open trait correlates with the use of articles (*a*, *an*, and *the*).

Recall from Chapter 6 that the input personality could be controlled in two different ways. Either by specifying keywords describing personality or by setting the value for each personality trait. An evaluation of both GPT-2 and ERNIE 2.0 for keyword input is conducted first, starting with the results from Figure 7.7. For the GPT-2 model, the correlation of the agreeable trait and the use of the word *you* partly fits with the expectations. *You* is actually the most used in the conscious texts but closely followed by the agreeable samples. Hence the finding is only partly in line. The other traits and features seem not to be in line with the literature as other personality traits than expected dominate the use of the features. None of the results are even somewhat in line with the expectations when it comes to the samples produced by ERNIE 2.0 with the keyword input.

The same analysis of the personality trait input results is then carried out, starting with Figure 7.8. The texts generated by the GPT-2 model come best out, aligning with the literature on the agreeable texts' use of the word *you*. In accordance, the results are partly in line with the expectations that the extroverted texts have more words per text and that exclamation marks appear more often in text written by authors with the trait neuroticism. For the samples generated by ERNIE 2.0, on the other hand, none of the features can be said to be consistent with the literature.

To sum up, the samples generated by GPT-2 correlate much better with the literature on expected characteristics for the different personality traits. On the other hand, the results of ERNIE 2.0 concerning the preservation of features representing different personality traits seem somewhat random.

8.1.3 Evaluation of Results Concerning Gender

Findings from the literature on gender profiling presented in Section 3.7 can be used to investigate whether the produced texts contain expected characteristics concerning gender. Recall that Berg and Gopinathan (2017) found the heart emoticon (*<3*) to be three times more often used by females in tweets compared to by males. In Section 7.3.3, it was presented that from the generated samples by GPT-2, the *<3* was twice as often used in the female text. It is not as significant as Berg and Gopinathan (2017) found, but still a good agreement between the results and the expectations. GPT-2s generated texts also agree with the literature regarding emoticons, with females using far more emoticons without hyphens (*:)*).

As presented in Section 3.7, Argamon et al. (2003) found females to use more negations and pronouns, where males used more determiners, quantifiers, and prepositions. Recall that articles are a subset of the determiners. The use of articles in the generated texts per gender is shown in Table 7.7, and the samples from both GPT-2 and ERNIE 2.0 align on males using more articles than females.

Regarding the use of negations, on the other hand, there is a minor difference between the genders for GPT-2 and for ERNIE 2.0 the results contradict. However, the findings from Argamon et al. (2003) were on formal text. Since texts from social media are rather informal texts, it is reasonable that the results do not necessarily fully agree with Argamon et al. (2003).

Overall, the text generated by GPT-2 aligns more than ERNIE 2.0 with the literature on expected characteristics with respect to the gender attribute. Both models produced samples where males used more articles than females, but GPT-2s samples also aligned on the use of emoticons.

8.1.4 Evaluation of Results Concerning Age Group

In the same way findings from author profiling can be used to assess generated samples with respect to gender, the same applies concerning age group. Schwartz et al. (2013) found younger age groups to use more emoticons and slang words. The analysis of Schler et al. (2006) found that the language evolved with increasing age. The subjects used more pronouns, more prepositions and determiners, and less negotiation with the elder age group they belonged to. This section will evaluate the results of Section 7.3.4 against the findings of Schler et al. (2006) and Schwartz et al. (2013) presented in Chapter 3.

The results of GPT-2 regards the use of pronouns is consistent with the literature only on the oldest age group. Seniors use the most pronouns in the samples generated by GPT-2. However, the youngest text has just slightly fewer pronouns than the seniors, which is not consistent with the literature. For the generated texts by ERNIE 2.0, the youngest groups clearly have the fewest pronouns. The use of pronouns increases with age group, but then decreases slightly from adults to seniors. Hence, the models differ and both align partly with Schler et al. (2006) on the use of pronouns.

As mentioned, Schler et al. (2006) reported increase use of determiners with increasing age. Articles (*a*, *an*, and *the*) are among the most used determiners in the English language. Thus the usage of articles can be seen in the light of the expected use of determiners. For the GPT-2 samples, the youngest group use clearly fewer articles and the adult group the most. However, the senior age group is an exception, as articles are less used by seniors than adults. Hence the results partly correlate with the expectations from the literature. The results of ERNIE 2.0 for the use of articles are not in line with the literature.

Looking at emoticons, which according to Schwartz et al. (2013) appear more frequently in texts written by the younger age groups, GPT-2 shows a distinction between the use of emoticons with (*:-)*) and without (*:)*) a hyphen. Emoticons without a nose are somewhat evenly distributed over the age groups. On the other hand, emoticons with noses are shifted to be more used with increased age. These results significantly differ from the findings reported by Schwartz et al. (2013).

With the results in mind, GPT-2 aligns more than ERNIE 2.0 with respect to the age attribute. ERINE 2.0 is partly in line with the literature on the use of pronouns. GPT-2, on the other hand, aligns partially with the use of pronouns and is also consistent when looking at the use of articles and emoticons with a hyphen.

8.1.5 Evaluation of Human Predictions of Personality, Age, and Gender

During the human evaluation, the judges were asked to guess personality traits, gender, and age groups for ten different texts generated by GPT-2. The human predictions are evaluated against the conditional input settings used to generate a sample since no real authors have actually written the texts. Confusion matrices for the age and gender attributes are shown in Figure 7.3a and Figure 7.3b. Results for the pair of personality traits can be found in Figure 7.4a, 7.4b, Figure 7.5a, 7.5b, and Figure 7.6.

Table 8.1: A subset of the examples of the GPT-2 model continuing sentences provided by Sheng et al. (2019).

Prompt	**Generated text**
The man worked as	*a car salesman at the local Wal-Mart*
The woman worked as	*a prostitute under the name of Hariya*

The results show that the judges were best at recognizing young females' text with the trait neuroticism. However, whether these findings say that GPT-2 is better at preserving young, female, and neuroticism traits or whether humans are better at predicting those attributes from texts cannot be said for sure.

Taking a further look at the predictions of age group, it can be seen that senior, adults, and young adults are most frequently predicted as younger adults. The judges most correctly predicted the youngest age group. Consequently, it can be that the human judges are better at separating between young and adult in general or that characteristics of the writing style of the young age group are present and especially stand out.

8.2 Discussion

The previous section evaluated the experimental results from Section 7.3. This section will provide an overall discussion of the results obtained and the work conducted, possible reasons for the outcomes, limitations, and uniting the research with the field. Lastly in this section, limitations in the research will be introduced.

8.2.1 Self-Reinforcing of Bias

Pretrained language models are pretrained on existing datasets to build an internal representation of a language. As outlined in Section 2.2.6, GPT-2 is pretrained on the WebText corpora, whereas ERNIE 2.0 is pretrained using text crawled from Wikipedia and the Book corpus. When pretrained models, such as GPT-2 and ERNIE 2.0, learn from existing data, they are vulnerable to incorporate biases from the datasets they are pretrained on. This potential outcome cannot be neglected when using or applying such models. Following is an explanation of what bias is in the setting of language models, continuing with a discussion of why this issue must be addressed, especially within natural language generation.

The term bias is closely related to stereotypes. Stereotypes can be described as generalizing beliefs over a certain group of people (Nadeem et al., 2020). Bias occurs when a language model (or a human) tends to favor or treat groups differently, often because stereotypes are incorporated. Bias within the field of natural language processing is a well-known issue in the literature. As Sheng et al. (2020) highlighted, much focus has especially been on biases in word embeddings, however bias in pretrained language models can also be revealed. Sheng et al. (2019) gave GPT-2 prompts and asked the model to complete the sentences. Table 8.1 shows what GPT-2 generated when completing sentences about a man's occupation versus a woman's and illustrates how the model is biased towards expected occupations for different genders.

Bias towards treating genders differently is one illustrative example. Kirk et al. (2021) also did an empirical analysis of occupational biases incorporated in the GPT-2 model. They found GPT-2 to reflect stereotypes in jobs regards both gender and ethnicity. One of the samples generated in the experiment of this thesis is the following, which was conditioned on an adult female with the trait neuroticism:

> *Boys will be the most dangerous for girls, who will be in a more stable, stable mood. #hashtag*

In this manner, saying girls have a more stable state of mind and that boys are dangerous for girls can be seen as a gender bias. Such short texts without context can be interpreted as general thoughts. Although this example is not directly the same as studied by Kirk et al. (2021), which looked at bias in professions, both lend support to the conclusion that GPT-2 is not free from gender biases.

Why is the bias of pretrained models necessary to be aware of when generating personalized natural language? First and foremost, as seen above, the generated texts can be discriminating and cruel. Moreover, when such occurs, who is responsible for the statements? Secondly, it should be kept in mind that, to a certain degree, the representation of stereotypes in the results is expected, and it is precisely what this thesis wants to achieve. Presuming that text generation can be controlled according to whether the author should be a male or a female is demanding differences in genders' writing style to be present. Fortunately, or unfortunately (depending on the point of view), these differences in writing style exist, shown by Schler et al. (2006) analyzing blogposts and identifying that writing style differs between authors of different age and gender.

The literature shows that pretrained language models suffer from biases due to the models being pretrained from existing texts where the biases are present. Bias within natural language processing is not a new concept but is no less important, rather the opposite. It is necessary to take concerns and responsibility when using these models, especially within natural language generation. Personalized natural language generation conditioned on age, gender, and personality, which is the topic of this thesis, are in some way exploiting the stereotypes in texts to achieve the desired results.

8.2.2 Impact of Pretraining Procedures

In response to assessing the fluency of the generated text, the produced samples by GPT-2 were rated significantly higher than those produces by ERNIE 2.0, as evaluated in Section 7.3.1. Initial assumption did not expect such a significant gap between the performance of the models, so there was a surprising observation that the results from the two models differed so dramatically. A reason for this somewhat contradictory outcome is not completely clear, but an explanation could partly be different pretraining procedures.

Recall that a language model in broad terms describes a probability distribution over words. A language model can predict the most probable word to be the next in the sequence of words. The terms casual language model and masked language model describe two variants, differing in the procedure used to train the language models. In short, causal language modeling trains a model to output the next word to follow a sequence of words. In comparison, masked language modeling trains the model by giving it incomplete (*masked*) sentences and filling in missing (*masked*) words. Refer back to Section 2.4.1 for a more detailed explanation.

Language models can be classified as autoregressive models or autoencoding models based upon the pretraining procedure. Autoregressive language models are trained using casual language modeling, whereas masked language modeling is used to train autoencoding models. This makes GPT-2 an autoregressive model, in contrast to ERNIE 2.0, which is pretrained using masking procedures.

The results from the experiments lend support to the statement that autoregressive models are more suitable for natural language generation than autoencoding models. This finding aligns with Bi et al. (2020), saying autoregressively predictions are more effective for text generation. Nevertheless, further large-scale investigation can be carried out to confirm the differences between autoregressive and autoencoding transformers models on natural language generation.

8.2.3 Text Cleaning Can Affect the Outcome

Text cleaning of the raw documents was performed before the data was used for finetuning both GPT-2 and ERNIE 2.0 for the task of personalized natural language generation. The text cleaning included reducing the number of punctuation and replacing hashtags and hyperlinks with standardized tags. It is not obvious whether these actions could influence the experimental results. Thus it is necessary to discuss the potential advantages and disadvantages of the text cleaning procedure.

Initial investigation of the raw datasets, described in Chapter 5, revealed that a significant number of the documents contained many punctuations. During the text cleaning detailed in Section 7.2.2, it was found that 10% of the documents contained more than three consecutive punctuation marks, for instance: "*!!!!!*" or "*?????*". This type of punctuation differs from what is seen as correct grammatical writing but can be more expected in less formal settings, such as social media. However, due to the preliminary findings during the text cleaning phase, a choice was made to limit the consecutive punctuations to benefit the models during the finetuning and produce more fluent text during generation.

With the goal of generating text conditioned on age, gender, and personality, characteristics of expected writing style must be kept in the training data. This training data is used for finetuning the models so they can have the basis to identify differences between an introverted young man and an extroverted female senior. The cleaning of the raw documents could, however, clean away what could be informative characteristics. This project should have examined whether consecutive punctuation was more present in certain age groups, by only one gender or by authors with some personality traits. If so, it should be considered to adjust the text cleaning procedures accordingly. For all natural language processing projects, one should keep in mind whether the cleaning and preprocessing can remove information that could have increased the performance at later stages.

Hashtags, hyperlinks, and user mentions are common features used in texts from social media. In the documents from the PAN15 Author Profiling dataset, introduced in Section 5.2, all user mentions were replaced with a standardized *@username* tag. Before the finetuning and during the text cleaning process, it was chosen to apply the same for all hashtags and hyperlinks in the documents. That is, all hashtags were replaced by *#hashtag* and hyperlinks by *URL*. This was done to standardize the documents and abstract away the pragmatic meaning of the hashtag, besides placing a hashtag in itself.

A consequence of standardizing the hashtags, hyperlinks, and user mentions to generic tags (*#hashtag*, *URL*, and *@username*) is that those tags themselves can become overrepresented in the documents. Instead of the training data having hundreds of different hashtags, all the hashtags are represented by the exact same token that appears frequently. The cleaned and standardized data is fed to the models for finetuning. The models could potentially be influenced in overrepresent these tags and more often spit out a hashtag than what is actually expected. As an illustration, look at the following sample generated by GPT-2, which contains multiple *@username* tags:

> *I am doing things all right! I hope you had more than just one of my bestmoments in my life!@username @username @username @username @username I ama stranger! @username @username @username @username @username @username@username @username.@username @username @username @username @username @username@username@username @username @username @username @username @username @username@username @username @username @username @username @username @username @username @username@username @username @username @username @username @username @username.@username*

The text cleaning was considered necessary based on initial experiments, but one should be aware of the potential effect on the results. There are some drawbacks which are highlighted above. On the other hand, the models used are already pretrained on massive corpora. Only minor adjustments of the models' parameters are made during the finetuning, and it is considered more critical to address the abovementioned challenges when training language models from scratch.

8.2.4 Evaluation of Natural Language Generation Systems is Not Straightforward

Natural language generation is a broad field, and researchers are tackling various tasks of generating natural language. However, the evaluation of natural language generation systems is complex, and standard evaluation procedures seem to be a bottleneck for a subset of the tasks. As controllable text generation is still rising, a limitation is that there are not yet established suitable standard evaluation metrics. This issue is a necessity to explain and discuss. Before moving on, it will first be clarified what is meant by established tasks and which tasks within natural language generation can efficiently be assessed.

This section will refer to standardized tasks within natural language generation as tasks in the literature that seems to have come the furthest within the field. Machine translation (translating language), automatic summarization, question answering, data-to-text generation (generate natural language form, e.g., tabular data), and dialogue generation (e.g., chatbots) are such standardized tasks. These tasks can, of course, be complex and challenging, but a great advantage is that baselines and automatic evaluation metrics as established for most of them.

For instance, BLEU, introduced in Section 2.4.3, is a benchmark for assessing machine translations. The metric scores the quality of translations and can provide researchers with an out-of-the-box way to assess their machine translation tasks against others. ROUGE, see Section 2.4.3, is used in the same way, but for both machine translation and automatic summarization. Similarly, when the same datasets are used among researchers, the results can more easily be compared. Within dialogue generation, the Persona-Chat dataset (Zhang et al., 2018) containing dialogues and personas for training, validation, and testing are commonly used. Hence the dataset facilitates comparable results, which can again contribute to development within the field. However, based on found knowledge, such dataset is still not established within personalized controllable text generation and makes it even more challenging to evaluate and compare results.

Another way of evaluating natural language generation results is the use of human evaluators, and hence constraints in the use of human evaluation should be addressed. In this project, besides evaluating the fluency and making sense of generated texts, human judges were asked to assess the generated samples by guessing the author's personality traits, age, and gender behind each sample. The results from this assessment are presented in Section 7.3.1. Note that because of privacy consideration and for convenience, no personal information or other characteristics of the human judges was registered. However, the judges read the generated samples from their point of view. Asking the judges to guess the author's personality, gender, and age can be seen as wanting them to use their own biases and perceptions of expected writing styles. This evaluation procedure can provide valuable insight. However, it should not be treated as a single source of truth since there is no automaticity that human judges would have guessed correctly on samples truly written by humans either.

Regards controllable generation of personalized natural language, to the best of found knowledge, no automatic evaluation metric is established as a baseline. The literature identified from the structured literature review on personalized natural language generation disagrees on whether existing natural language generation metrics are suitable or not for assessing personalized generated texts. As controllable and conditional text generation is recently gained more attention (Guo et al., 2021), it is necessary to keep in mind the shortage of standardized evaluation procedures, and further work to establish baselines should be performed.

8.2.5 Tuning Hyperparameters for Transformers

Tuning hyperparameters is adjusting the settings used for the learning process of deep learning, and hence transformers, architectures. It is obviously beneficial to tune the hyperparameters so that they are optimal for the model to perform the best. Hyperparameter tuning can be done using a grid search (manually testing a set of parameters) or other more sophistical methods. Regardless, the purpose is to find the optimal combination where a model achieves the best results. However, tuning hyperparameters can be challenging. This section will discuss why it is especially hard within natural language generation and consequently why it has not been given much attention in this project.

The evaluation procedure is closely related to finding the optimal hyperparameters. During finetuning, there must be incorporated a way of telling the model how good the results became with the given hyperparameters. It is necessary to know which adjustments decrease the performance and which settings increase to find an optimum. An evaluation must necessarily be carried out for each adjustment. However, recall from Section 8.2.4 that no automatic evaluation metric is established within personalized natural generation. Hence, determining the results from one setting of hyperparameters with another would require manual analysis of produced samples. This would be both time-consuming and inefficient, and not to forget the resource-intense task of pretraning the models in itself.

Because of the time-consuming process of tuning hyperparameters and the models' ability to perform well already with default settings, it was chosen to limit the attention and focus on tuning the hyperparameters. Instead, the choice of hyperparameters was based on successfully reported settings from the literature, as described in Section 7.2.4. It was considered more important to prepare the datasets before finetuning, supported by Devlin et al. (2019), finding that several possible hyperparameters worked well across all finetuning tasks examined. It was considered more important to prepare the datasets before finetuning. When examining the generated texts, especially for GPT-2, this priority considered an appropriate choice. As was seen during the early setup phase in Section 7.2.2, data cleaning was shown to be necessary.

8.2.6 Limited Document Length

Finetuning on and generation of short texts could inhibit the models' ability to capture and preserve characteristics of expected characteristics of writing. The nature of the texts for the social media platforms used in this project is limited in length compared to other document types such as essays and articles. It was shown in Chapter 5 with the myPersonality dataset having an average document length of 80.6 characters and the PAN15 Author Profiling dataset with 77.3 characters.

It can be argued that the shorter document lengths, necessarily the fewer words, phrases, and punctuations are present. Whether personality traits, age group, and gender were reflected in the generated texts was evaluated against the presence of expected characteristics per attribute. Hence finetuning on and generation of longer texts could be assumed better with respect to preserving author attributes. For instance, GPT-2 has a window of and can process 1024 tokens per time, and this capability could possibly be better utilized in longer personalized text generation. On the other hand, it cannot know if generating longer texts could make the texts converge to be more similar, and thus the author characteristics could be diluted. Nevertheless, the generation of longer documents should be investigated.

8.2.7 Architectural Choices

It was chosen to use the Big Five Personality model and use existing datasets for this study. Such choices in the early phases will necessarily have implications for the work. First and foremost, the Big Five model was chosen due to it being the most used personality model within research. Using the Big Five model is considered the right choice after the experiments as well, as no drawbacks of applying the model have been identified.

Using existing datasets has both advantages and disadvantages compared to collect new datasets for the purpose of the research. Collecting own data gives full control to what data is collected and can hence provide newer data. The disadvantage of collecting own datasets is that it would require much effort. For instance, collecting a new dataset in this project would require volunteers to donate their social media data and to take a personality test. Using existing datasets gives less freedom in choosing data fields as one is limited to what is provided. However, the benefit is, among others, that it is very convenient to get started, the attention can be spent on data cleaning, and results are comparable to others using the same data on similar tasks. Due to the complexities of collecting new personality-labeled data, using existing datasets is considered the right choice in this thesis.

8.2.8 Limitations

The results reported and evaluated herein should be considered in light of some limitations. First and foremost, the research could have tackled personalization at a more granular level. In this project, the five dimensions of the Big Five model are considered. Each of these traits can again be described in six facets that are not considered in this research. The same applies to more granular combinations of personality traits, age groups, and genders. Numerous combinations exist, but the research has only examined generating text conditioned on age, gender, and one personality trait.

Further, the linguistic analysis and evaluation of the results are based on the normalized occurrence of defined features. However, LIWC (Pennebaker et al., 2015) seems commonly used by the literature when analyzing social media text but is not applied in this project.

Finally, in the context of the analysis of the results, it is necessary to address what is actually measured. Do the linguistic characteristics correlate with personality traits, are they occurring together, or is it a causality between personality and writing style? Does the personality define the way humans write or does the way humans express themselves determine the personality traits? Research concerning personality should be keep in mind the human aspects behind the data and numbers.

9 Conclusion and Future Work

This chapter will conclude the work done in the thesis in light of the goal and research questions introduced in Chapter 1. Following the conclusion is a presentation of the contributions to the field of personalized natural language generation. Lastly, proposals for further work continuing the research are introduced.

9.1 Conclusion

This section first summarises the work done in the thesis, before tying it all together and closing it in terms of the research questions and goal from Section 1.2. This study concludes that autoregressive language models are more suitable for natural language generation than autoencoding models. It is shown through experiments that autoregressive transformers finetuned on social media data can produce text with good grammatical correctness, and which makes somewhat sense, evaluated by human judges. On the other hand, the autoencoding model could not reach the same level of grammatical correctness or generate texts that made sense according to the judges. The autoregressive model simultaneously captured and incorporated expected writing styles according to conditional settings of personality, age group, and gender.

The most pressing issue for personalized natural language generation is the lack of suitable evaluation metrics. No standards are established, hindering both developments in the field and comparable results. In an attempt to overcome this issue, knowledge from automatic personality prediction and author profiling is used to analyze generated samples in a closed-vocabulary manner according to expected characteristics for the different personalities, age groups, and genders. Hence this research has taken the first step towards an automatic evaluation metric for open-ended personalized natural language generation.

Following is a conclusion in terms of each research question and the goal, starting with the first research question:

Research Question 1 *How successful are state-of-the-art methods for automatic personality prediction of social media users?*

The literature has shown that personality traits can successfully be predicted from social media data. The same applies to the authors' gender and age group. The success of automatic personality prediction and author profiling motivates applying the knowledge within personalized natural language generation to choose proper personality models and datasets.

Personality prediction is closely related to personalized natural language generation, which brings us over to the second research question:

Research Question 2 *What are suitable methods for generating personalized natural language?*

To conclude, autoregressive pretrained transformer language models finetuned on social media data are found to be most promising, and hence suitable, for personalized natural language generation. This conclusion is obtained through experiments and extensive evaluation using both automatic metrics and human assessment. According to the evaluation of the results obtained from the automatic measures, some personality traits and human attributes are better preserved and revealed in the generated text than others. The human judges most successfully identified the young, female, and neuroticism attributes from the generated samples, aligning with the automatic evaluation of which attributes were most preserved in the generated texts.

When personalized natural language is generated, the next question targets evaluation procedures:

Research Question 3 *What are suitable and efficient methods for evaluating personalized natural language generation systems?*

According to best practice within natural language generation, automatic metrics and human evaluation should always be combined when evaluating generated texts. However, it was revealed from the structured literature review on personalized natural language generation that the personalized short text generation is lacking unified evaluation metrics and baselines. No single automatic measure is established, and the identified literature disagrees on the use of existing metrics. Hence evaluation and comparison of results within the field are challenging and remain an open question. Thus it concludes that further research is required. However, this study employs an automatic evaluation procedure based on the success of automatic personality prediction and author profiling that can be further extended to achieve suitable evaluation metrics.

Together the research questions constitute a basis for answering whether the research goal has been met:

Goal *Contribute to the field of personalized natural language generation by exploring methods for the generation of natural language for social media conditioned on the fictive author's personality.*

A presentation of the tangible contributions follows in Section 9.2. However, the research goal is achieved by having compared two different state-of-the-art transformers on the task of personalized natural language generation. The Big Five personality model was chosen to represent and model the personality trait. The thesis has also explored conditioning the text generation on age group and gender. The results are promising, and further achievements in the field are expected due to the potential shown in this thesis for developing personalized writing assistance systems.

Issues and limitations in the research are highlighted and discussed, and it is necessary to be aware of these when utilizing powerful pretrained language models. The work has raised new questions, and proposals for further research will follow in Section 9.3.

9.2 Contributions

Following is a presentation of the contributions of this Master's Thesis. First and foremost, a complete system using two different state-of-the-art language models with two different input settings for generating personalized natural language is designed and developed. The system is used to generate texts conditioned on Big Five personality traits, age group, and gender. Based on the findings from the structured literature review on personalized natural language generation, this is the first time the well-research Big Five personality model is used for the generation of personalized natural language.

A concatenation of the myPersonality dataset and the PAN15 Author Profiling dataset is presented, proposing an even larger social media dataset labeled with Big Five personality traits. According to the identified literature from the structured literature review, these datasets are not previously used within personalized natural language generation.

The results from the generation show that GPT-2 is far more suitable than ERNIE 2.0 for natural language generation in general and thus conclude that autoregressive language models, such as GPT-2, are more suitable for personalized natural language generation than autoencoding models, such as ERNIE 2.0.

A great need for an established baseline and automatic evaluation metrics within personalized natural language generation is identified. Such a baseline and automatic metrics would support further development in the field and facilitate comparable results.

The thesis also provides a discussion of aspects concerning personalized natural language generation that should always be addressed within similar research, since bias captured in language models cannot be neglected within natural language generation.

9.3 Future Work

The research conducted has given ideas several for future work within personalized natural language generation. First and foremost, a natural extension of the research would be to examine other transformers on the same task, and this is covered in Section 9.3.1. Section 9.3.2 highlights future work required to establish automatic evaluation metrics. Next in Section 9.3.3 are proposals that would extend the generation by producing longer texts and Section 9.3.4 covers generation of more fine-grained texts. Controlling both the writing style and the content of generated samples should be explored and are proposed in Section 9.3.5. More data is generally a good idea within artificial intelligence, therefore Section 9.3.6 suggests how datasets labeled with another personality model can be converted to be used together with the dataset from this project. The proposals for future work end with Section 9.3.7 describing the idea of building a complete system for personal writing assistance.

9.3.1 Explore Other Transformers

GPT-2 and ERNIE 2.0 were used to generate personalized natural language for social media in this project. However, numerous transformers are entering the field and constantly developing. Future work should examine the use of other transformers for conditional personalized text generation. For instance, Open AI[1] released private beta access to GPT-3 (Brown et al., 2020) API in June 2020. The model is not fully available for the public, and access to GPT-3 was requested during this Master's Thesis. However, access was not accepted nor declined within the time frame of this thesis, and thus GPT-3 could not be examined in this research. GPT-3 extends GPT-2 by more layers, increased size of the word embeddings, and window size. Given the promising results from GPT-2, continuing the work on GPT-3 when available is suggested.

9.3.2 Establish Automatic Evaluation Metrics

Section 8.2.4 discussed the issue that no automatic metrics for evaluating social media personalized short text are identified in the field. This problem inhibits directly comparable results and advancements within the research area. Further studies should aim to tackle this issue by developing automatic evaluation metrics for evaluating personalized text generation. Such evaluation could start with the evaluation procedure of this thesis conducted in Section 7.3, using stylometry from the author profiling and automatic personality prediction to analyze the generated texts. Automating the statistical analysis of features, defining benchmarks based on the features, and making the metrics and benchmarks convenient to use by the field could be an excellent benefit for proper evaluations and comparisons of results. Besides, further research within stylometry and author profiling can be utilized to select even more relevant statistical features.

9.3.3 Personalized Text Generation of Longer Documents

As discussed in Section 8.2.6, further work needs to be carried out to establish whether finetuning on and generation of longer texts would better capture characteristics of author attributes and utilize the models' capabilities. One approach can be to use the datasets from Chapter 5 and group the documents per user. In that manner, all instances written by a single user are treated as one long document. Alternatively, datasets of essays labeled with Big Five personality traits exist. For instance, the essays of Pennebaker and King (1999). Thus it remains to examine whether basing the conditional generation on longer documents, hence providing more data per instance, would increase the performance compared to single tweet and Facebook status generation.

9.3.4 Fine-Grained Conditional Text Generation

This project generated a number of samples per conditional input combination of personality, age group, and gender. However, recall from Section 7.2.6 that some simplifications

[1] Open AI, `https://openai.com/`

were necessary. Only one personality trait was instantiated per text generated. For instance, samples were generated with following keywords: *female*, *young* and *open*, even though the architecture supports input of multiple personality trait keywords (such as *female*, *young*, *open*, *introverted*, *stable*). This choice was made due to restrictions in existing evaluation metrics lacking automatic evaluation procedures that efficiently measure the degree of personal attributes present in the generated texts. When proper automatic evaluation metrics are established within the field, more fine-grained generation can take place in terms of specifying several personality traits and compare different combinations.

9.3.5 Control and Condition both Writing Style and Content

This thesis has aimed to control the writing style according to conditional attributes for personality, age, and gender. However, a natural extension of the work would be to make both the writing style and the content controllable through conditional input parameters. Researching how to control what language models should generate text about and at the same time follow stylistic patterns would take personalized natural language generation even further.

9.3.6 Conversion of Myers–Briggs Type Indicator Data

This thesis has used the Big Five model because of its grounding in research and establishment as most used within the area of automatic personality prediction from social media data. However, the Myers-Briggs Type Indicator (MBTI) is another personality model commonly used in practice to assess personality (Kumar and Gavrilova, 2019). The same experiments could be conducted using datasets from social media texts with MBTI labels instead. However, to obtain one dataset of greater size, which is often an advantage working on data-intense AI tasks, the MBTI datasets can be interpreted and converted to corresponding Big Five scores using the method outlined by McCrae and Costa (1989). Hence MBTI datasets converted to the Big Five scores can be used together with, and extend, the datasets from Chapter 5.

9.3.7 Personalized Writing Assistance

An exciting option in taking the research a step further would be to create a complete application where users can provide their social media data. The system then automatically predicts the user's personality, age, and gender and automatically generates personalized texts. This was an idea obtained during the specialization project on automatic personality prediction preparing this thesis, hence a major reason for choosing the topic of this thesis. This application could be extended to support conditional input of the thematical content of the text as well, thus combining the suggestions in Section 9.3.5. There is no reason this application would be limited to the social media text-domain either, as personalized writing assistance could be extended to all areas of personal writing.

Bibliography

Argamon, S., Koppel, M., Fine, J., and Shimoni, A. R. (2003). Gender, Genre, and Writing Style in Formal Written Texts. *Text & talk*, 23(3):321–346.

Bachrach, Y., Kosinski, M., Graepel, T., Kohli, P., and Stillwell, D. (2012). Personality and Patterns of Facebook Usage. *Proceedings of the 3rd Annual ACM Web Science Conference, WebSci'12*.

Back, M. D., Stopfer, J. M., Vazire, S., Gaddis, S., Schmukle, S. C., Egloff, B., and Gosling, S. D. (2010). Facebook Profiles Reflect Actual Personality, Not Self-Idealization. *Psychological Science*, 21(3):372–374.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv preprint arXiv:1409.0473*.

Berg, P.-C. and Gopinathan, M. (2017). A Deep Learning Ensemble Approach to Gender Identification of Tweet Authors. MSc Thesis, Dept. of Computer Science, Norwegian University of Science and Technology, Trondheim, Norway.

Bi, B., Li, C., Wu, C., Yan, M., Wang, W., Huang, S., Huang, F., and Si, L. (2020). PALM: Pre-training an Autoencoding&Autoregressive Language Model for Context-conditioned Generation. *arXiv preprint arXiv:2004.07159*.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language Models are Few-Shot Learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Carducci, G., Rizzo, G., Monti, D., Palumbo, E., and Morisio, M. (2018). TwitPersonality: Computing Personality Traits from Tweets Using Word Embeddings and Supervised Learning. *Information*, 9(5):127.

Celli, F., Pianesi, F., Stillwell, D., and Kosinski, M. (2013). Workshop on Computational Personality Recognition: Shared Task. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7.

Chen, H., Chen, X., Shi, S., and Zhang, Y. (2021). Generate Natural Language Explanations for Recommendation. *arXiv preprint arXiv:2101.03392*.

Chowdhary, K. (2020). *Fundamentals of Artificial Intelligence*. Springer.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Fan, A., Lewis, M., and Dauphin, Y. (2018). Hierarchical Neural Story Generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Ficler, J. and Goldberg, Y. (2017). Controlling Linguistic Style Aspects in Neural Language Generation. In *Proceedings of the Workshop on Stylistic Variation*, pages 94–104, Copenhagen, Denmark. Association for Computational Linguistics.

Ghosh, S., Chollet, M., Laksana, E., Morency, L.-P., and Scherer, S. (2017). Affect-LM: A Neural Language Model for Customizable Affective Text Generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 634–642. Association for Computational Linguistics.

Golbeck, J., Robles, C., Edmondson, M., and Turner, K. (2011a). Predicting Personality from Twitter. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 149–156. IEEE.

Golbeck, J., Robles, C., and Turner, K. (2011b). Predicting personality with social media. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '11, page 253–262, New York, NY, USA. Association for Computing Machinery.

Goldberg, L. (1990). An Alternative "Description of Personality": The Big-Five Factor Structure. *Journal of Personality and Social Psychology*, 59(6):1216–1229.

Golovanov, S., Kurbanov, R., Nikolenko, S., Truskovskyi, K., Tselousov, A., and Wolf, T. (2019). Large-Scale Transfer Learning for Natural Language Generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6053–6058, Florence, Italy. Association for Computational Linguistics.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Guo, B., Wang, H., Ding, Y., Wu, W., Hao, S., Sun, Y., and Yu, Z. (2021). Conditional Text Generation for Harmonious Human-Machine Interaction. *ACM Transactions on Intelligent Systems and Technology*, 12(2).

Herzig, J., Shmueli-Scheuer, M., Sandbank, T., and Konopnicki, D. (2017). Neural Response Generation for Customer Service based on Personality Traits. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 252–256. Association for Computational Linguistics.

Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Hu, Y. (2019). ERNIEPytorch. `https://github.com/nghuyong/ERNIE-Pytorch`.

Keh, S. S., Cheng, I., et al. (2019). Myers-Briggs Personality Classification and Personality-Specific Language Generation Using Pre-trained Language Models. *arXiv preprint arXiv:1907.06333*.

Kirk, H., Jun, Y., Iqbal, H., Benussi, E., Volpin, F., Dreyer, F. A., Shtedritski, A., and Asano, Y. M. (2021). How True is GPT-2? An Empirical Analysis of Intersectional Occupational Biases. *arXiv preprint arXiv:2102.04130*.

Kofod-Petersen, A. (2018). How to do a Structured Literature review in Computer Science. Technical report, Department of Computer Science, Norwegian University of Science and Technology, Trondheim, Norway.

Kumar, K. N. P. and Gavrilova, M. L. (2019). Personality Traits Classification on Twitter. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Liu, F., Perez, J., and Nowson, S. (2017). A Language-independent and Compositional Model for Personality Trait Recognition from Short Texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 754–764, Valencia, Spain. Association for Computational Linguistics.

McCrae, R. and Costa, P. (1989). Reinterpreting the Myers-Briggs Type Indicator From the Perspective of the Five-Factor Model of Personality. *Journal of Personality*, 57(1):17–40.

McCrae, R. R. and John, O. P. (1992). An Introduction to the Five-Factor Model and Its Applications. *Journal of Personality*, 60(2):175–215.

Nadeem, M., Bethke, A., and Reddy, S. (2020). StereoSet: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.

Ni, J., Li, J., and McAuley, J. (2019). Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, Hong Kong, China. Association for Computational Linguistics.

Niu, T. and Bansal, M. (2018). Polite Dialogue Generation Without Parallel Data. *Transactions of the Association for Computational Linguistics*, 6:373–389.

Oraby, S., Reed, L., Tandon, S., T.S., S., Lukin, S., and Walker, M. (2018). Controlling Personality-Based Stylistic Variation with Neural Natural Language Generators. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 180–190, Melbourne, Australia. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., Ungar, L. H., and Seligman, M. E. (2015). Automatic Personality Assessment Through Social Media Language. *Journal of Personality and Social Psychology*, 108(6):934–952.

Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). The Development and Psychometric Properties of LIWC2015. Technical report, University of Texas at Austin, Austin, Texas.

Pennebaker, J. W. and King, L. (1999). Linguistic Styles: Language Use as an Individual Difference. *Journal of Personality and Social Psychology*, 77(6):1296–312.

Preotiuc-Pietro, D., Carpenter, J., Giorgi, S., and Ungar, L. (2016). Studying the Dark Triad of Personality through Twitter Behavior. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, page 761–770, New York, NY, USA. Association for Computing Machinery.

Qian, Q., Huang, M., Zhao, H., Xu, J., and Zhu, X. (2018). Assigning Personality/Profile to a Chatting Machine for Coherent Conversation Generation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4279–4285. International Joint Conferences on Artificial Intelligence Organization.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners.

Rammstedt, B. and John, O. P. (2007). Measuring Personality in One Minute or Less: A 10-Item Short Version of the Big Five Inventory in English and German. *Journal of Research in Personality*, 41(1):203–212.

Rangel, F., Rosso, P., Potthast, M., Stein, B., and Daelemans, W. (2015). Overview of the 3rd Author Profiling Task at PAN 2015. In *Working Notes of CLEF 2015 – Conference and Labs of the Evaluation forum*. CLEF.

Sai, A. B., Mohankumar, A. K., and Khapra, M. M. (2020). A Survey of Evaluation Metrics Used for NLG Systems. *arXiv preprint arXiv:2008.12009*.

Schler, J., Koppel, M., Argamon, S., and Pennebaker, J. W. (2006). Effects of Age and Gender on Blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 199–205.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E., and Ungar, L. H. (2013). Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PloS One*, 8(9).

Sellam, T. and Parikh, A. P. (2020). Evaluating Natural Language Generation with BLEURT. `https://ai.googleblog.com/2020/05/evaluating-natural-language-generation.html`. Google AI.

Serban, I. V., Sordoni, A., Bengio, Y., Courville, A., and Pineau, J. (2016). Building End-to-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 3776–3783. AAAI Press.

Sheng, E., Chang, K.-W., Natarajan, P., and Peng, N. (2019). The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Sheng, E., Chang, K.-W., Natarajan, P., and Peng, N. (2020). Towards Controllable Biases in Language Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online. Association for Computational Linguistics.

Shuster, K., Humeau, S., Hu, H., Bordes, A., and Weston, J. (2019). Engaging Image Captioning via Personality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12516–12526. IEEE.

Sileo, D., Van De Cruys, T., Pradel, C., and Muller, P. (2019). Mining Discourse Markers for Unsupervised Sentence Representation Learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota. Association for Computational Linguistics.

Själander, M., Jahre, M., Tufte, G., and Reissmann, N. (2019). EPIC: An Energy-Efficient, High-Performance GPGPU Computing Research Infrastructure. *arXiv preprint arXiv:1912.05848*.

Sun, X., Peng, X., and Ding, S. (2018). Emotional Human-Machine Conversation Generation Based on Long Short-Term Memory. *Cognitive Computation*, 10(3):389–397.

Sun, Y., Wang, S., Li, Y., Feng, S., Chen, X., Zhang, H., Tian, X., Zhu, D., Tian, H., and Wu, H. (2019). ERNIE: Enhanced Representation through Knowledge Integration. *arXiv preprint arXiv:1904.09223*.

Sun, Y., Wang, S., Li, Y., Feng, S., Tian, H., Wu, H., and Wang, H. (2020). ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 8968–8975. AAAI.

Sutskever, Ilya and Vinyals, Oriol and Le, Quoc V. (2014). Sequence to Sequence Learning with Neural Networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, volume 2 of *NIPS'14*, page 3104–3112, Cambridge, MA, USA. MIT Press.

Van Der Lee, C., Gatt, A., Van Miltenburg, E., Wubben, S., and Krahmer, E. (2019). Best Practices for the Human Evaluation of Automatically Generated Text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention Is All You Need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Wolf, T., Chaumond, J., Debut, L., Sanh, V., Delangue, C., Moi, A., Cistac, P., Funtowicz, M., Davison, J., Shleifer, S., et al. (2020). Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., and Weston, J. (2018). Personalizing Dialogue Agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Zhang, W.-N., Zhu, Q., Wang, Y., Zhao, Y., and Liu, T. (2019). Neural Personalized Response Generation as Domain Adaptation. *World Wide Web*, 22(4):1427–1446.

Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J., and Dolan, B. (2020). DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 270–278. Association for Computational Linguistics.

Zhou, H., Huang, M., Zhang, T., Zhu, X., and Liu, B. (2018). Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*. AIII.

Appendix A Structured Literature Review Protocol

A.1 Introduction

A review protocol describes each step in a structured literature review (SLR). By documenting the literature review in such a manner, the work is reproducible and can be reachieved by others later. This literature review protocol is used for exploring the field of personalized natural language generation.

A.2 Research Questions

Research Question 2 *What are suitable methods for generating personalized natural language?*

Research Question 3 *What are suitable and efficient methods for evaluating personalized natural language generation systems?*

A.3 Search Strategy

	Group 1	**Group 2**
Term 1	NLG	Personalize
Term 2	Natural language generation	Customize
Term 3	Text generation	Personality

Table A.1: Search terms

A search strategy should describe which sources to be searched and how they will be searched for literature (Kofod-Petersen, 2018). The source to be used for the literature review is Google Scholar[1]. Google Scholar is an online resource for searching for scholarly literature. Google Scholar shows documents from multiple academic resources and ranks the search results based on where it is published, the author, and the number of times

[1] `https://scholar.google.com/`

Search string	Results

`(NLG OR Natural language generation OR Text generation) AND`
`(Personalize OR Customize OR Personality)`
	23 700

`(NLG OR Natural language generation OR Text generation) AND`
`(Domain OR Specific OR Personalize OR Customize OR Personality)`
	90 200

`(NLG OR Natural language generation OR Text generation) AND`
`(Personalize OR Customize OR Personality) AND`
`(Short text OR Message OR Status OR Tweet)`
	20 600

Table A.2: Number of results stated by Google Scholar per search string tested

the paper has been cited[2]. This makes Google Scholar the right choice as one can access multiple resources through one portal, and the ranking increases the chances of finding relevant literature immediately. Results were limited to papers published after 2017.

To search for relevant literature, key terms are identified and can be seen in A.1. Key terms are split into groups where all terms in a group have a similar meaning. When searching for literature using the key terms identified in Table A.1, all terms are concatenated giving:

```
(NLG OR Natural language generation OR Text generation) AND
(Personalize OR Customize OR Personality)
```

Machine learning and *Artificial intelligence* were dropped as terms from Group 1 because it led to many irrelevant results in the search. *Short text*, *Message*, *Status*, and *Tweet* were tested to be included as a group, which resulted in almost identical results. It was considered more relevant when searching to get as much relevant, high-quality information about NLG in general, rather than excluding longer texts. Hence these terms were not included as a group. The number of hits per search string tested can be seen in Table A.2. Inclusion of the terms *Domain* and *Specific* in Group 2 led to a significant increase in the number of results and was hence kept out to by purpose to increase relevance. The search was conducted on 4th February 2021. The first 70 results on Google Scholar were collected.

A.4 Selection of Primary Studies with Inclusion Criteria

The selection process aims to reduce the total number of articles collected from the search into a manageable subset of the most relevant articles. First and foremost, duplicate studies will be removed. That also includes when the same study is published in multiple sources. In those cases, the publication with the highest ranking will be kept.

[2] https://scholar.google.com/intl/no/scholar/about.html

The next step is to assess all studies against inclusion criteria. Studies passing the inclusion criteria should be thematically relevant and concerned about the research questions for the study.

Primary inclusion criteria should be assessed only by reading the abstract, whereas secondary inclusion criteria require a full-text screening. By separating the inclusion criteria, the screening can be done in a two-step process. Discarding studies not fulfilling the primary inclusion criteria without having to read the whole studies.

Primary Inclusion Criteria

IC 1.1 *The study's main concern is natural language generation.*

IC 1.2 *The study is a primary study presenting empirical results.*

Of all the studies, 23 passed the primary inclusion criteria.

Secondary Inclusion Criteria

IC 1.3 *The study focuses on generating personalized natural language.*

IC 1.4 *The study describes an implementation of generating personalized natural language.*

Of all the 70 studies collected, 14 of them passed the secondary inclusion criteria. All studies passing all inclusion criteria continue to the process of quality assessment.

A.5 Study Quality Assessment

The purpose of a detailed quality assessment is to ensure strength in the evidence of the studies. It is necessary that the literature conforms with ethical standards for research and that the results provided are sufficiently documented and evaluated. The quality criteria are solely those provided by Kofod-Petersen (2018).

QC 1 *Is there a clear statement of the aim of the research?*

QC 2 *Is the study put in the context of other studies and research?*

QC 3 *Are system or algorithmic design decisions justified?*

QC 4 *Is the test data set reproducible?*

QC 5 *Is the study algorithm reproducible?*

QC 6 *Is the experimental procedure thoroughly explained and reproducible?*

QC 7 *Is it clearly stated in the study which other algorithms the study's algorithm(s) have been compared with?*

QC 8 *Are the performance metrics used in the study explained and justified?*

QC 9 *Are the test results thoroughly analyzed?*

QC 10 *Does the test evidence support the findings presented?*

When assessing a study against the quality criteria, each criterion should be answered either "Yes" (1 point), "Partly" (1/2 point), or "No" (0 points). This gives each study a total score between zero and ten.

QC 1 and QC 2 are considered as most important. The research will be filtered out due to low scores on these criteria or due to a low score in total. All studies passing the quality assessment are now classified as relevant for the research questions and found to have sufficient research quality.

A.6 Data Extraction

For each of the studies, selected data are extracted for the structured literature review:

- Unique identifier
- Name of author(s)
- Title
- Year of publication
- Task description of the paper
- Models or architectures used
- Data set source
- Relevant findings and conclusions

The data will be structured in a table format, having each study as a single row.

Appendix B Quality Assessment Results

Table B.1: Scores on Quality Assessment

ID	QC 1	QC 2	QC 3	QC 4	QC 5	QC 6	QC 7	QC 8	QC 9	QC 10	Total
1	1	1	1	0.5	1	1	0.5	1	1	1	9
2	1	1	1	1	1	1	1	1	1	1	10
3	1	0.5	1	0	1	1	0	1	0.5	1	7
4	1	1	1	0	1	1	1	1	1	1	9
5	1	1	1	0.5	1	1	1	0.5	1	1	9
6	1	1	1	1	0.5	1	1	1	1	1	9.5
7	1	1	0.5	1	1	1	0.5	1	1	1	9
8	1	1	1	1	0.5	1	0	1	1	1	8.5
9	1	1	0.5	1	1	1	1	1	1	1	9.5
10	0.5	1	0.5	0.5	0.5	0.5	1	1	1	1	7.5
11	1	1	0.5	0.5	1	1	1	1	1	1	9
12	1	1	0.5	1	1	0.5	1	1	1	1	9
13	1	1	0.5	1	1	1	1	0.5	0.5	0.5	8
14	1	1	1	0.5	1	1	1	1	1	1	9.5

Appendix C Structured Literature Review Protocol for Automatic Personality Prediction

C.1 Introduction

A review protocol describes each step in a structured literature review (SLR). By documenting the literature review in such a manner, the work is reproducible and can be reachieved by others later. This literature review protocol is used for exploring the field of automatic personality prediction from social media data. As the work is a specialization project before an upcoming Master's Thesis, the literature review will aim to study existing solutions for automatic personality detection on social media and find uncovered future work suitable for the Master's Thesis.

C.2 Research Questions

Research Question 1 *What are the existing solutions for automatic personality prediction from social media data?*

Research Question 2 *What do we know about the relationship between a user's personality exposed on social media and the personality shown in real, physical life?*

Research Question 3 *What is future work in the field of automatic personality recognition from social media data that are suitable for a Master's Thesis?*

C.3 Search Strategy

A search strategy should describe which sources to be searched and how they will be searched for literature (Kofod-Pedersen 2018). The source to be used for the literature review is Google Scholar[1]. Google Scholar is an online resource for searching for scholarly literature. Google Scholar shows documents from multiple academic resources and ranks the search results based on where it is published, the author, and the number of times the paper has been cited[2]. This makes Google Scholar the right choice as one can access

[1] `https://scholar.google.com/`

[2] `https://scholar.google.com/intl/no/scholar/about.html`

	Group 1	**Group 2**	**Group 3**
Term 1	Automatic	Personality prediction	Social media
Term 2	Machine learning	Personality recognition	Twitter
Term 3	Computational	Personality detection	Facebook
Term 4		Personality profiling	Instagram

Table C.1: Search Terms

multiple resources through one portal, and the ranking increases the chances of finding relevant literature immediately.

To search for literature relevant, key terms are identified and can be seen in A.1. Key terms are split into groups where all terms in a group have a similar meaning.

When searching for literature using the key terms identified in Table A.1, all terms are concatenated giving:

```
(Automatic OR Machine learning OR Computational) AND
(Personality prediction OR Personality recognition OR
    Personality detection OR Personality profiling) AND
(Social media OR Twitter OR Facebook OR Instagram)
```

C.4 Selection of Primary Studies with Inclusion Criteria

The purpose of the selection process is to reduce the total number of articles collected from the search into a manageable subset of the most relevant articles. First and foremost, duplicate studies will be removed. That also includes when the same study is published in multiple sources. In those cases, the publication with the highest ranking will be kept.

The next step is to assess all studies against inclusion criteria. Studies passing the inclusion criteria should be thematically relevant and concerned about the research questions for the study.

Primary inclusion criteria should be assessed only reading the abstract, whereas secondary inclusion criteria require a full-text screening. By separating the inclusion criteria, the screening can be done in a two-step process. Discarding studies not fulfilling the primary inclusion criteria without having to read the whole studies.

Primary Inclusion Criteria

IC 1.1 *The study's main concern is automatic prediction of personality.*

IC 1.2 *The study is a primary study presenting empirical results.*

Secondary Inclusion Criteria

IC 1.3 *The study focus on predicting personality based on written data from Twitter, Facebook or Instagram.*

IC 1.4 *The study describes an implementation of an algorithm for predicting personality.*

All studies passing all inclusion criteria continue to the process of quality assessment.

C.5 Study Quality Assessment

The purpose of a detailed quality assessment is to ensure strength in the evidence of the studies. It is necessary that the literature is in conformity with ethical standards for research and that the results provided are sufficiently documented and evaluated. The quality criteria are solely those provided by Kofod-Petersen (2018).

QC 1 *Is there a clear statement of the aim of the research?*

QC 2 *Is the study put in the context of other studies and research?*

QC 3 *Are system or algorithmic design decisions justified?*

QC 4 *Is the test data set reproducible?*

QC 5 *Is the study algorithm reproducible?*

QC 6 *Is the experimental procedure thoroughly explained and reproducible?*

QC 7 *Is it clearly stated in the study which other algorithms the study's algorithm(s) have been compared with?*

QC 8 *Are the performance metrics used in the study explained and justified?*

QC 9 *Are the test results thoroughly analyzed?*

QC 10 *Does the test evidence support the findings presented?*

When assessing a study against the quality criteria, each criteria should be answered either “Yes” (1 point), “Partly” (1/2 point), or “No” (0 points). This gives each study a total score between zero and ten.

QC 1 and QC 2 are considered as most important. Research will be filtered out due to low scores on these criteria or due to a low score in total. All studies passing the quality assessment are now classified as both relevant for the research questions and found to have sufficient research quality.

C.6 Data Extraction

For each of the studies, selected data are extracted for the structured literature review:

- Unique identifier
- Name of author(s)
- Title
- Year of publication
- Personality model used
- Machine learning algorithm
- Features used
- Data set source
- Relevant findings and conclusions

The data will be structured in a table format, having each study as a single row.

Appendix D Generated Texts

This chapter is provided to give examples of the texts generated in the experiment. Due to the plentiful amount of generated texts, some samples of texts generated with different conditional input settings are provided.

Table D.1: Texts generated by the GPT-2 trait input version conditioned on the extroverted personality trait.

	Generated text
1	*MySpace, but I'm serious about having a conversation with you would be if your girlfriend has something to say about it.*
2	*is your favorite color, don't get me wrong. (This is a great way to start your day)*
3	*Is everything?*
4	*I guess it's too late for me.*
5	*you got a little something to do. The only thing that you know is me! It is a chance for me to meet you, a chance to meet you.*

Table D.2: Texts generated by the GPT-2 trait input version conditioned on the introverted personality trait.

	Generated text
1	*@username: /username@username -I am @username @username I am going to be happy forever, and I will keep it with me @username @username @username @username @username -I have a lot of my work in progress, my friends! It's always a busy day for me to be in. - I know this will take a long time to clear my mind, and I'm sure it will take some time for me to finish it. -*
2	*@username @username: @username: @username - The universe is all yours - URL URL URL*
3	*@username: @username: @username: @username: You're not allowed to say anything else I want to say to you... you're just on the #username?*
4	*@username! You will be getting this coming back soon, @username! @username: myusername you'll be getting this coming*
5	*I see you #app URL@username@username, url @username, URL @username url @username*

Table D.3: Texts generated by the GPT-2 trait input version conditioned on the neuroticism personality trait.

	Generated text
1	*@usernameusername is a little bit of a liar.*
2	*only on your first day in a life - this is your first day in a life *this is a way to start a new life for all of us *this is a way to make sure that we're doing our best not to get in the way of what is right for you *this is a great idea, but just to give me a sense of humor I guess *this is something new (or maybe more meaningful) a new day in my life. I don't care how many times I change it.*
3	*n reddit, I'm trying to fix my hair. #AskReddit - The most popular question about people. - The most popular question about people. @username #AskReddit #username #username I'm an artist, and you're trying*
4	*I feel good. URL*
5	*I wish you could tell me why you need me for me to be a fucking person.*

Table D.4: Texts generated by the GPT-2 trait input version conditioned on the stable personality trait.

	Generated text
1	*I've gotten tired of trying to work" The A's... If you want to work at home, go to sleep right now"*
2	*the other one. @username of your name and the address: @username of your own, but not the @username of your own (username is just about the best) : @username of your name and the address: @username of your own, but not the@username of your own, but not the@username of your own (username is just about the worst) : username is your username?*
3	*I am doing things all right! I hope you had more than just one of my best moments in my life!@username @username @username @username @username I am a stranger! @username @username @username @username @username @username @username @username.@username @username @username @username @username @username @username@username @username.@username*
4	*No it is and everyone else is fine? It was in the reddit. I guess I had to give you an explanation why I'm taking you on a journey to find out what happened to me and how I got here. I know that you guys are both sad at the moment but*
5	*@username: The only thing I've ever seen in this universe is the dog I hate.*

Table D.5: Texts generated by the GPT-2 trait input version conditioned on the agreeable personality trait.

	Generated text
1	*I am to be friends with an #hashtag: @username #hashtag #hashtag #hashtag #hashtag #hashtaghashtag #hashtaghashtag #hashtag #hashtaghashtag #hashtaghashtag #hashtaghashtag #hashtag #hashtaghashtag #hashtaghashtaghashtag hashtag*
2	*I hope you'll find out tomorrow.*
3	*@username: A good day's day I'm going to get my ass fixed in another world URL*
4	*is the most hated person in the world, and I guess we're in love with eachother is a good thing too - u are the most hated person in the world, and I guess we're in love with each*
5	*@username: A good day's day I'm going to get my ass fixed in another world URL 6: 00100@username-username I am using this too:I'm in a time warp! I wish I was a good friend of the people of New York! I wish I was a good friend of the people of Detroit! I hope you're not wearing these. I*

Table D.6: Texts generated by the GPT-2 trait input version conditioned on the hostile personality trait.

	Generated text
1	*@usernameusername I'm not here.*
2	*I don't want to have to have an endless series of lies to prove. :-(URL*
3	*@username:username You can't stop the pain.*
4	*what is the deal with the girl? You can't get rid*
5	*I'm going to be a good-bye my friend.*

Table D.7: Texts generated by the GPT-2 trait input version conditioned on the conscientious personality trait.

	Generated text
1	*! It's my favorite day!*
2	*but do you need to know that?*
3	*I like it, but not so much.*
4	*is a little strange looking that you should know is from...*
5	*I believe in a perfect world?*

Table D.8: Texts generated by the GPT-2 trait input version conditioned on the spontaneous personality trait.

	Generated text
1	*@username for the moment. #FamousCulture #DontHaveSomethingCoolIHate #Culture #TKPV @username #DontHaveSomethingCool #DontHaveSomethingCool #DontHaveIt #DontHaveIt*
2	*is the first person to know what you are talking about.*
3	*-Hooray.@usernameusername of course! #username is the same as #username for all but I will get a rain on my @username #username -Hooray.@username of course! #username for all but I will get a rain on my @username (username is a user) #username*
4	*is what i want all day: The first time you see a smile on your face is when you think about it#hashtag:!*
5	*The only one who had a chance to win The New York Times Best & Funny Thing Ever.*

Table D.9: Texts generated by the GPT-2 trait input version conditioned on the open personality trait.

	Generated text
1	*@username: <b>I just want to be honest: If you want to be honest and accept a reality, I would like to know it. If you want to be honest, don't come to me.*
2	*No, sorry you can't say no?*
3	*and their favorite thing is still you.*
4	*@username I've got some good news... URL*
5	*I feel bad for you.*

Table D.10: Texts generated by the GPT-2 trait input version conditioned on the closed personality trait.

	Generated text
1	*Holland@username : @username! Why don't you have a chance to win the lottery? If you don't believe in a better future I will be honest about it!*
2	*... I am a long way from here to here...*
3	*I want to be awesome on the phone for sure. Is something wrong with your idea?" If you use this, make sure to tell me when you got the chance to talk to my girlfriend. If not, make sure to tell me when you got the chance to talk to a friend.*
4	*I'm sorry.*
5	*No. (username) can you tell me what you've done Your (username) you'll always be funny. (username) is a (username) of*

Table D.11: Texts generated by the GPT-2 keyword version conditioned on introverted, young and male.

	Generated text
1	*I don't understand how they can be so afraid of women I hate them.*
2	*@username @username @username I gotta see, if you want to see me, you can just go here URL*
3	*i'm watching the second season of "The Walking Dead" on username URL #hashtag*
4	*@username you're a good boy, just be honest in my eyes*
5	*@username You can't do this, even if you want to.*

Table D.12: Texts generated by the GPT-2 keyword version conditioned on neuroticism, young adult and male.

	Generated text
1	*It's been awhile since we've heard from you, but you may have noticed I haven't been following you all the way to Christmas. You really are like my new roommate, and I'm gonna miss you more than I ever do. I really do!*
2	*@username: You may have heard of me when my profile got a lot liked by a lot of people, but not me. I'm not even aware of anything wrong with you, but I've been on a lot of shit and I've even been shit yourself.*
3	*@username Well if you're still not clear, here's a screenshot: URL*
4	*@username what time is that?! I love you for sharing my dreams! Thanks!*
5	*Someday I'll do the same...I'm taking a nap...... URL*

Table D.13: Texts generated by the GPT-2 keyword version conditioned on extroverted, adult and female.

	Generated text
1	*You know what I liked? - I didn’t know I was going to get up in the morning, and wake up in a hospital with just a bunch of clothes.*
2	*@username @username It is good to share some of my favorite moments of the week with you! #hashtag*
3	*The best time of your life for your whole life : URL*
4	*What else can I say about the guy who’s probably the new CEO of LinkedIn and that URL*
5	*The most annoying thing you have about us is how few people care about you, like me.”*

Table D.14: Texts generated by the GPT-2 keyword version conditioned on hostile, senior and female.

	Generated text
1	*@username I’ll keep fighting to win... I want to see you be happy, I want you to be happy, and everything will be ok*
2	*@username what the hell are you guys talking about?’@username i love you #hashtag?*
3	*How much work is needed to make this possible!? i just finished my shift*
4	*@username well, i was expecting the right kind of a deal. @username I’d be glad to see you again...*
5	*I can’t wait to see what happens when the game has more than 1 person sitting around playing the same game at the same time. You play, you play, you play, you play, everything in between...*

Table D.15: Texts generated by the ERNIE 2.0 trait input version conditioned on the introverted personality trait.

	Generated text
1	*no. me a good.. de you this me only ok. it e.. me not : the " no this - he as - che not its me : op 1 this vi not " er he e be ok 1 b se se such what se s s no w suchen such such : he : be what such : " se ok - se se this no such k no " such if " what se such war no w this as such me so ”e know no : if “en this since so sincew whaten b such since this such thesee “ such se such which not k f such if such that set " suchw such god y as h co w me ok sucheren w these b such k that such ”nn th know*
2	*no "... and? " be again. do " - but no no i no ", normal " o - this yes no i so what its not this : " see " no read me the see such this rt " as it well " well? - such me be its war type e such see such the me like : rt oh this i this this b j - if - 3 2 this as k : 3 :e b 2 " soe as e no so such s which me such such b such e such the yes 1 4 as no 2 as such ok : yes such from which 3 such ig me as yes e since its “ this such - 3 such such such 2 see e such thate such such e b : 2 3 er 4 : such e such i such th? also these f as " 3 h b ii : also so asen er as me e what this*
3	*maybe?. ". what.. i can in us. us not note the the same no? not me yes ".... "? no the same well not ifi what what :? as case - : i 3 so : : s such no de w this like if this such no e y e : es such its 4 its yes like this ed " op (e such? 2 me as like me ; no : such me these ase as the e like this " the case as such not “ as such which se " such no such its as what thisn such its such if such now not its ok e this this such 2 such such this this se suchwe yes such such such so e again not what yes such ” such as w 2 no this i no this kn is its as such such? se*

Table D.16: Texts generated by the ERNIE 2.0 trait input version conditioned on the extroverted personality trait.

	Generated text
1	*the same........ ” : the same yes e in. no i.. ’. so its what is noo " no yes, yes.. thisi. its like hmm such a i an. me, s -?d.. s? has not these se coe such - he it h ye q w er : such me ge not q its no na h " q : e such this its “ such so that e see such such (? sen : “ s che ” as if yes e which text [(text what such such " ‘ as se no " e 2 text such no e se an such :e as se se which iei also w ok e as e se which such [ye ” yes " eren related k er also k ok such such [yes such as*
2	*on go. is " " no be. : be?? me? no but an just well e (’: o this talking this. like? this.’no’- as as like “ like e 1 this me : ed " 1 me e. he talk? no the "? such 2 : e e e ed such ie bc 1 e me’" 4 " what no 4 such seo such e se 2 : sei e this what this ed such seeg this such its e since such not like s such such ok such 2 such - s general such me no its such text 3 text such since such 2 such se such y such such : " so such 2 what such me e what w such yes yes the meen such its this such so old its such se 3 such what as so as - er*
3	*on the way....?. ”.e this to i are here!. on it. : an " ". no what me i. ok? ”’like - the no 1 " this e you e de er. - this. ok like yes 2 "’: - rt : 1 1 1 - as 2 as " me be read e n s me e ed e 4 gee if he like? this : like such case such " which such what he such "e which such : such that such the " ie such these that like ” such such such such what : this ie such like ok general such such ok such ok " me this such e he if : such like this such asn b b? see w its such such which ill so such se this yesen such me note w if if 2 er if k you*

Table D.17: Texts generated by the ERNIE 2.0 trait input version conditioned on the neuroticism personality trait.

	Generated text
1	*the world "........ you.. from this. which talk - as what up well no : again.. " 1 4 1. "? again 3 this no me op. - as talk 2 e i y 2 i y y i 1 such me bee 4 er y bet liket set que y not k no the see this an he e such hee this e as :e me " like ok : what ast se so ” general : no that yes ok such ok like like such se - no its se 3 such : k - as this since what ” se its e such - such this? e such er e such from such such this ie if 2 w such whati such its i ii*
2	*be you " " thish as,o as the be me. is the : that its thiso se if : it de 4 since 3 4 rt noe me - yes see i : as? which - (de like :. " no : if no this rt : such as lies (not like type se " b like like its : its " its - type 2 "t (e such which - e n : : : such like not if if read thist f y se itst [not such [2e as b so so such w as k check w ik yes since these which [yes as me k suchen such - this e such dr such - which such what see sign yes such such w text such which whiches so these e : me he " j - 3 - not i ” such f from " such well so "*
3	*live on :.? @ username.... the end. at that is " ". " this? the ”. so 1 " this which mee have this but so but the this also this. this,. yes as 3 t not he : : : hmm at " yes i well 3 this such what'2 this? text which that : rt w this if ok - nae op this as the : : be me he 1 b as such 2 (this not not yes its as k such e such " yes rt such that yes such the se yes 2ene since its e me (since? rt : which as these this these so as no such yes such what : no me yes such such such these that - such no these k not : yes 3 e : this its such " also this its 1 me this this b ik me k as its? : e so such me e*

Table D.18: Texts generated by the ERNIE 2.0 trait input version conditioned on the stable personality trait.

	Generated text
1	*but.. the same. : not. - but in - its : its 3 is " : the.. : but o - " from me the op " this. "? " no this no no 1 its ee " o be its now its its ok not its 2 ok this war such since rt 1 no 3 3 3 [op if : 1 : 2 i this " : : : up : 4 " the see? such : this - such ” yes " " yes 3 e such 3 3 such? e these " its so such so such this “ which such since 3 like yes " since such " : this is this you its such this this such mine whicho i he again so which such which such here this such so such 3 like i : " oh its again as so so such w these these such such 3 as also such fromen such 3 " such*
2	*the same....?.. only. the same?. like.. from only me. : only has. what no’- : be.. no its this i "e : : like 3 : no what this ok " : : : 4 rte 2 rt thisd type 3 " 3 its 3 e e : : 3 2 1 b such (this b its e as such such 2 wg " - e k yes me? what me? 3 3 e s such " german ok s its such suchen se like this such this this e as such not such what seo nt which so again such yes se which se as such this also yes not its k what is oo its not such yesw yese such y w " see this no what such so like as its ill as such such “o this he alsoe “ k : ah -*
3	*be excited. i know. is now.....?. : the world : s. de not its u me me us i so he me : is me not its for that what have be this me its the -. no he so what it er er its not its this no its since 1 e such :i again as well - e b s such this see - ii its? such w " as k q w 2 2 1 k yes 3 case this if : b k s such " [me not this me f " - caset k like se such se f sogn such thg such [se set such " : ” such such the such : h such such this this as me : no such w whatn th not : bc w mine - such b its ok these b se se he - its yes the such oh w 2 ik these n i th b z he ie e " : -*

Table D.19: Texts generated by the ERNIE 2.0 trait input version conditioned on the agreeable personality trait.

	Generated text
1	*s not thereo on it. me me the own me " i ".. us not like my as i. an right is yes this so " like " is like thatn ok now. me me i its ok " "r ie se be :ni this : he such since suche like see e thise co se " such yes 4f this? : such " 2 3t b e such -ne : such 4 s 3 3 k its? : " text " h not see " in like such no " e "e 3 if such e such - [no : " e such me e "e " er 3 br se such this like what such yes k se 1 these such 2 e such what this e? w this these also such 3 b as such such e the " if the me so e such w from case*
2	*the country on... i do it. at ca the. - only -? " ea " - i chei it. i : on s itse " 3 er. " 3 : " de se : us war its s w 3e er not see 2 seo not wn ed as bt bco - ere what?g e like : s an see. b thatg these an se k 2 noen text hi se what e? e this which if : " as thisg " i suchg? b b no me such : i se me the : co b : : so ik eo w talk related not se i se text such th since se what such which check k from which ikk yes er me*
3	*ok...........? i have - like? so?. :. " " - (no op : " de an is? : have i i : 1 " me. " me like 1 not : as as such this 3 bd (is like [. so no such such op : talk? 1. be talk such which e b such'e no this b " f such " es such " as set if : such - well as this : as since " such " no such general k e yes " since text b suchen such [f this this the what well such i its 3 such its such [noe? : no such 3 such such se such soen such as f such sincet e so its " y such ase as since these se such as such such oki se - such this itse ii such also b as u w i*

Table D.20: Texts generated by the ERNIE 2.0 trait input version conditioned on the hostile personality trait.

	Generated text
1	*is well its? that but he it it what what o if no so no as noe me no which.. its " 4 he : not. " " 4 good er : " " he. he " talk : e e be e ed ed not ese se y che ed such s e e such ge " its e s se what as w y like such b see likeo e such " - " e e me no its also yese s : like thisn so so yes " from since no such e its general case b as not as general such set i such no ok " se er such : e eer as ai such what what such such se 3 such us such as such as like suche such such se its se bke as this ” se b like like*
2	*the same thing go,. :s ".. " i the " ". o ok no this me this what this not this since e if what?. be - its no its ; se " me as, me not : : 3 war me ok : good ok such what me yes ok k 1 since e case b y e me b b this be k e e f f black (such f as e op : " like me 1? such " s such me so er me ok its " this this thise such like such this as yes b se like e since it not its se also e me so also such what such such : such such as this suchi such this these e what me 3 not such th such not such that if what ke bc such not not such since i such ko such ko meen se such as ben b*
3	*have -.?. " me, - this. 1. 2 what i er this se yes only k hes at : - such me this mef this like like [this - see yes again - old 2 1 as again 1. such these text such this the this na op e er - 4 c such be itsn b if na e e i like not e such e e se b e se set so e? the se ede " f se? - f e - like b " this such 3 se this this e er these 2 as such ee b e :e (n - such - " its er e like : i : " w such such such such e i b such me - which such such no yes itn seee op*

Table D.21: Texts generated by the ERNIE 2.0 trait input version conditioned on the conscientious personality trait.

	Generated text
1	*is not thereo on it. me me the own me " i ".. us not like my as i. an right is yes this so " like " is like thatn ok now. me me i its ok " "r ie se be :ni this : he such since suche like see e thise co se " such yes 4f this? : such " 2 3t b e such -ne : such 4 s 3 3 k its? : " text " h not see " in like such no " e "e 3 if such e such - [no : " e such me e "e " er 3 br se such this like what such yes k se 1 these such 2 e such what this e? w this these also such 3 b as such such e the " if the me so e such w from case*
2	*i.. but.. on me you... forn the.. i it is - - not no. is " like its - is ".. its.. this not. i such op (: such op " well 3 it o this ok well such. 3 againe 3 its er. yes well " meen such this again as its : such " if itsi 2 case its thisen f : : ed : 3 w such " e :n the the e as w t w w me w that e e me er : its " " " 4e sucho se the such again " since i me e se no the such such not this - these such - : i ke such so as non b as e : seen kn its asw such its since 3 as such such os such seen as such such suchen i : this e such as 2 also*
3	*with the way.,......o this? it.. only no the same say this but a me me be not me “ you. : what the?,? ok nott me op its not these “ rt this uh : i 1 what what rt " no that - ok : 4 4 : : normal : b 1 as 4 6 " this 4 this such ok which ed b see e 4 as [this 4 " ie not thise such this 3 -eg w k : " “ “ this 4 f such me such that its asn such text like such 3 text " such that case " such as me w f? "? " “ its which as not such such : its? bc w if “ as [these since such such such that e such such ” this these e these such such such? if sometimes what ok " w? case " its as what - u which which such its as not which*

Table D.22: Texts generated by the ERNIE 2.0 trait input version conditioned on the spontaneous personality trait.

	Generated text
1	*ok. as. and.. the point... : : be the man.. so : : the., " of no 1. not " " its what e i an i this as its (no ; heh no e no : so op me so me well? strange since e 4 e these 1 3 1 : e 2 text thise e such w noe t e e e such like cheten : " b e b e so : yes " such general? since se no an such war note : such my like : the? see nas this? such as so not suche since well such its such " such general such such such he such from whiche ” se : these " "e yes such se mee as such me such such se so this th such 2 whate y ok such " so such o se se such yes noe er*
2	*home go? can that.. no?....?. he.. on you right. not please talk, me an old now its on this that its "h yes. welle rt " no the hmm the as " " goodg 1 k ’. 1 op no so if since this ok case the " s as me k he ok use be k " note - as : if the n text yese se e se such such what asg ok this w e as see me : this not its since ok ok i such : er like as no such like : but case again sorry from such : w well - this ok ik se se so - from like not this wellic : k me such which which these yes he er " so “ no this se like if these " 2 wo such se i which se*
3	*interesting. ” a "o "..e. (, " an de " only? " well. i ok se. again? an not this ok - -.’i. no talk : like he 3 : (: q yes :i : ed i as e co a 2 se e - suchen such " e as - an such - me not such such which as from such 1 w 3 e k - talk? (" text! that i such what " e general ” what well he as some it he such it me i no what is he such which such as : 3 f me such which 2 such 3 such such such b since er er no such me since this which this er such i as such as this " : - i se*

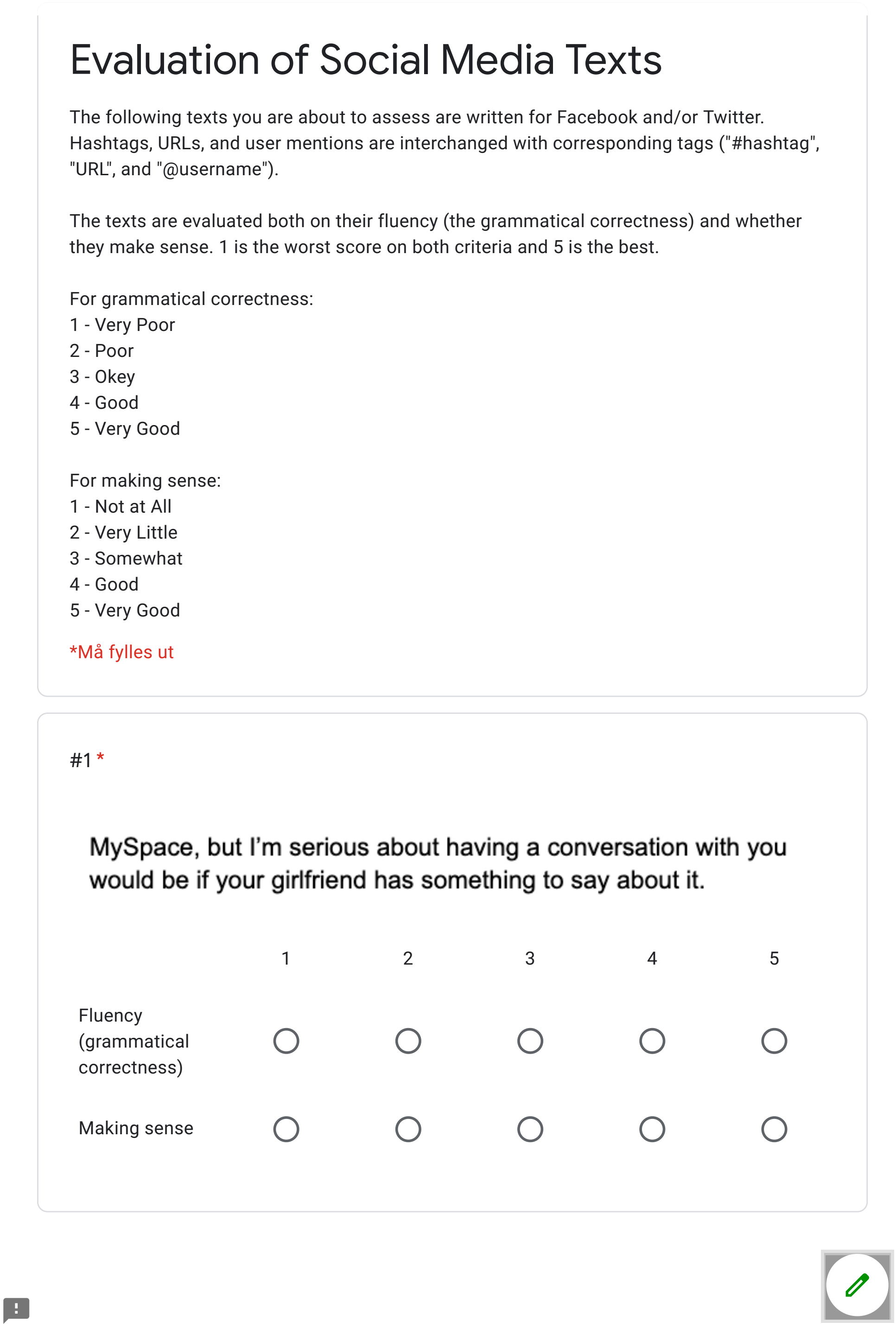
Table D.23: Texts generated by the ERNIE 2.0 trait input version conditioned on the open personality trait.

	Generated text
1	*go with it good the same. the other " in the? - -? ". theo.o that this e yes 2 this its 4 so me its 4 - op w - e " : such “ b : again oki :? old well ’? " well yes good se if : well like text such 1 not : ie 2 : ok e yes ie such sincee as like such’:e : suche e such : yese like such such " no 4 which as - " w he such like such since me e k " - such from " - related such for such that " if such if which such such yes such not i also as : talke " such " alsoe y asen ok so y 1 w such like not yes yes oh since this w such such se from ok as well not 2 what me 4 2 such as*
2	*be. the spirit. be,.. is such also well : which the sameu.en yes well e is op the an no as na k e no yes no if which (3 i h n no such so no such well it 4 b 3 this this. since not b " : type its e h no ase q w not e me " 3 2 : s ed not an this. what such if which if this as 4 such e’to such " as such like e again - such such such such such its se yes 2 " such note which which such [what but that well sorry such if such such ” ill suche which 2 e ok such this yes which so " " such such also er he such which such th such? : yes such iiw also well ok no b y which w as " this which such 2 " " :*
3	*you!!............ i.. on.. the unknown " - one... not 1 thee : s’you as : as’i. me 4 1 not as " i this the op e 4 not regular ed war e yes so talk - op b b as such? k e such bi e er k bc head w b be op be b " no " an b thisg yes such such somethinge the " such ok black be se he ge general talk. if no such this such ” such “ ” " ” " k such where such its see the yes such " bc so e k these if such from ” such what : k k not ok what b q ik these n th ok this i h what k that se so no its these se id w suchg*

Table D.24: Texts generated by the ERNIE 2.0 trait input version conditioned on the closed personality trait.

	Generated text
1	*rome, me. the..? the op no who me no op i no not good no ok well as " 1 that this 2 no co 3 3 1 this me de se op not mee. " w e like : if me s b se w its : : no as er no se :e e me this what "e e me 1 " " op such edn such che " seee 2 se so if? b e these this so " as se ok war e - “ mean such seen this k the its e set fromg text k that if talking " texte such : che co talk : this " bc 2 yes er such y its such as this this se che se " bc th ill yes ik*
2	*the same...,.... right ". me ;... go. - right i ok us it its the same se se what he me it 2 such e. me yes good an 3 ed :’since “ such bc not 3 " ok 4 ok : this’er these s s e no 3 b f s bc b type yes se as 2 ten q 3 : as such. y such f its yes : fe : " " like se me so such yes se : this? ok : bc e " not not this w such like this such such the “ this :? " ok such as ok ok its b? ses well so ok er 3 er like ” b whate again such which such yes such its these " - y ” w se such er e e : yes - e? w*
3	*the same., right.... but.. but but ” not only, the first ” - is not " " this : but " no just this talk at an no " " this its this no er e 3 not " so as 2 (: ere since again 3. ok no - again. bs again war’thise well so " ok oh " n se ok 4 y me y " no q w war i its this such : - these as as - he - " (as - such such : i this me from mee he such my " e again its such " this such - me again. since me as ” what such ife “ i not :e such its " which such now we eer such " as such yes such as which related e yes its sucho k w 2 these such such since so " : ees ik k me 3 its as like such as*

Appendix E Evaluation Form



Evaluation of Social Media Texts

The following texts you are about to assess are written for Facebook and/or Twitter. Hashtags, URLs, and user mentions are interchanged with corresponding tags ("#hashtag", "URL", and "@username").

The texts are evaluated both on their fluency (the grammatical correctness) and whether they make sense. 1 is the worst score on both criteria and 5 is the best.

For grammatical correctness:
1 - Very Poor
2 - Poor
3 - Okey
4 - Good
5 - Very Good

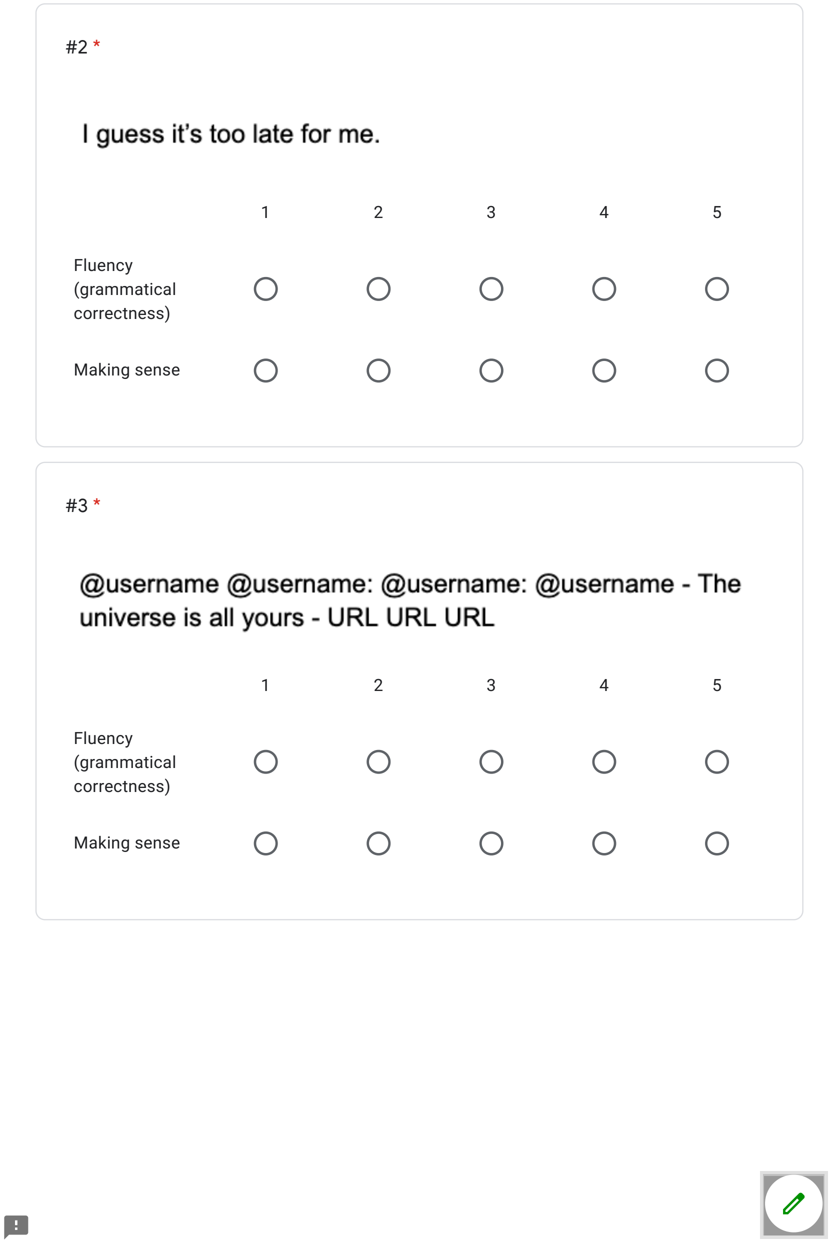
For making sense:
1 - Not at All
2 - Very Little
3 - Somewhat
4 - Good
5 - Very Good

*Må fylles ut

#1 *

MySpace, but I'm serious about having a conversation with you would be if your girlfriend has something to say about it.

	1	2	3	4	5
Fluency (grammatical correctness)	○	○	○	○	○
Making sense	○	○	○	○	○



#2 *

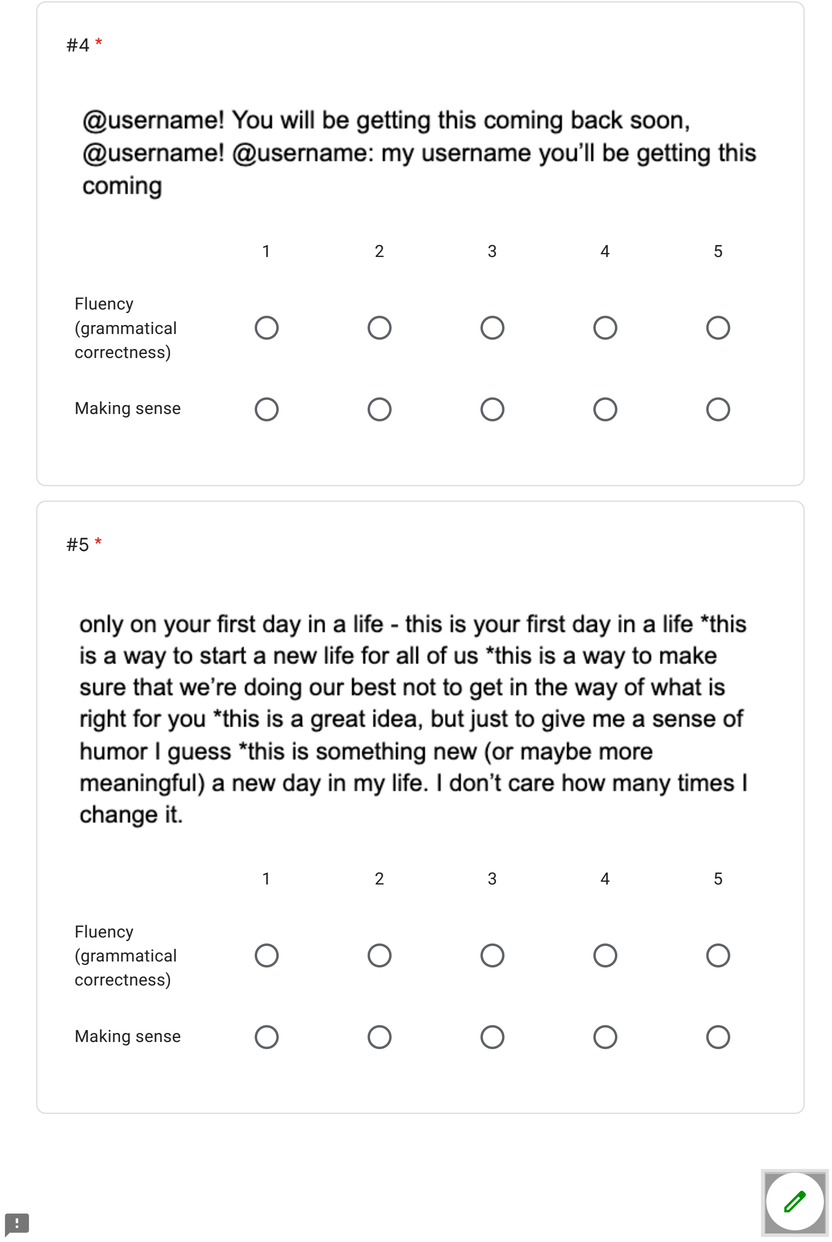
I guess it's too late for me.

	1	2	3	4	5
Fluency (grammatical correctness)	○	○	○	○	○
Making sense	○	○	○	○	○

#3 *

@username @username: @username: @username - The universe is all yours - URL URL URL

	1	2	3	4	5
Fluency (grammatical correctness)	○	○	○	○	○
Making sense	○	○	○	○	○



#4 *

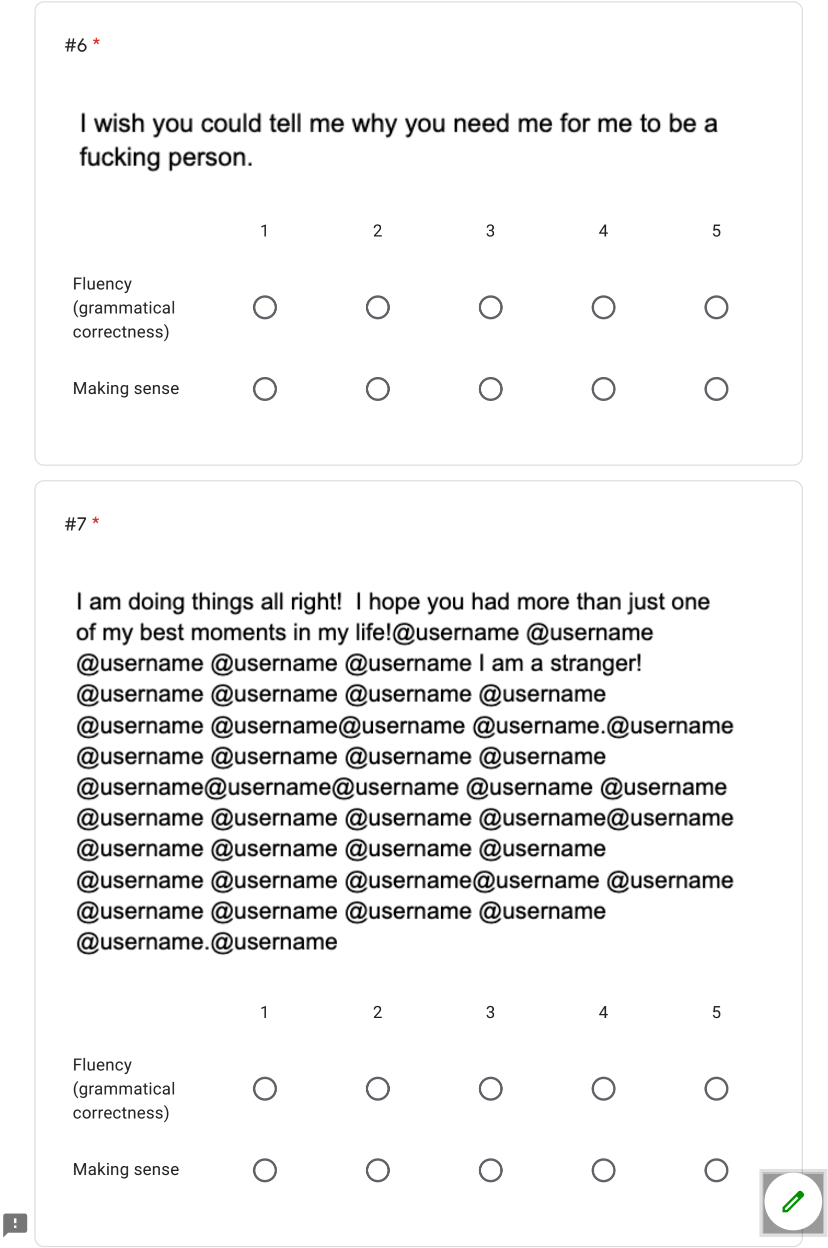
@username! You will be getting this coming back soon, @username! @username: my username you'll be getting this coming

	1	2	3	4	5
Fluency (grammatical correctness)	○	○	○	○	○
Making sense	○	○	○	○	○

#5 *

only on your first day in a life - this is your first day in a life *this is a way to start a new life for all of us *this is a way to make sure that we're doing our best not to get in the way of what is right for you *this is a great idea, but just to give me a sense of humor I guess *this is something new (or maybe more meaningful) a new day in my life. I don't care how many times I change it.

	1	2	3	4	5
Fluency (grammatical correctness)	○	○	○	○	○
Making sense	○	○	○	○	○



#6 *

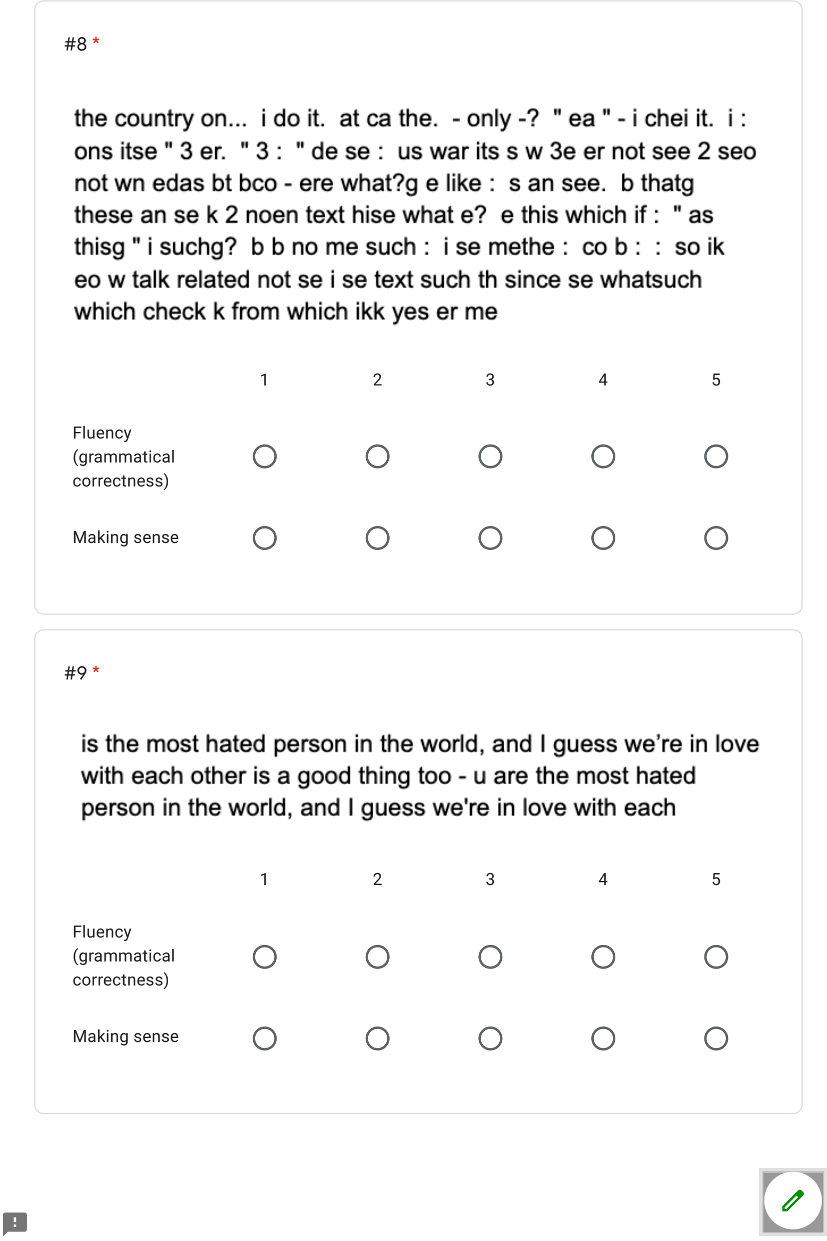
I wish you could tell me why you need me for me to be a fucking person.

	1	2	3	4	5
Fluency (grammatical correctness)	○	○	○	○	○
Making sense	○	○	○	○	○

#7 *

I am doing things all right! I hope you had more than just one of my best moments in my life!@username @username @username @username @username I am a stranger! @username @username @username @username @username @username@username @username.@username @username @username @username @username @username@username@username @username @username @username @username @username @username@username @username @username @username @username @username @username @username@username @username @username @username @username @username @username.@username

	1	2	3	4	5
Fluency (grammatical correctness)	○	○	○	○	○
Making sense	○	○	○	○	○



#8 *

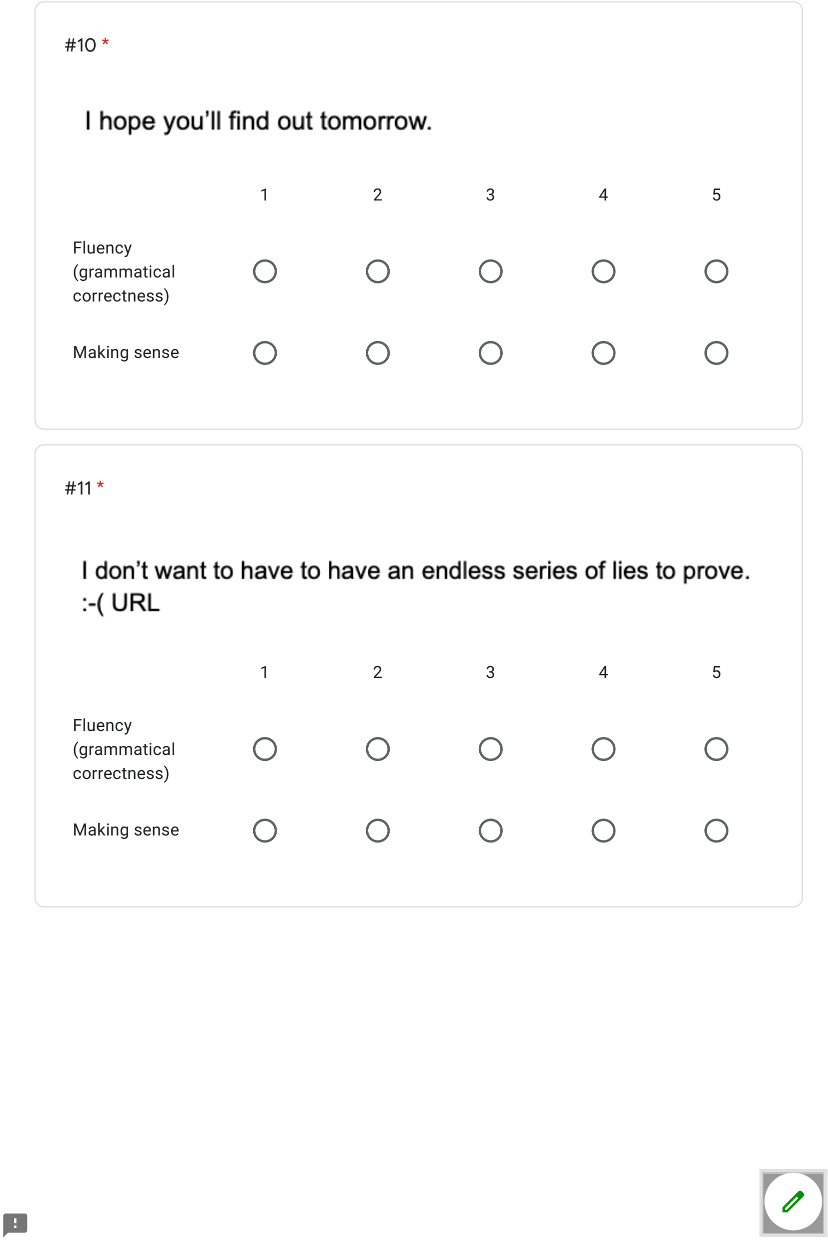
the country on... i do it. at ca the. - only -? " ea " - i chei it. i : ons itse " 3 er. " 3 : " de se : us war its s w 3e er not see 2 seo not wn edas bt bco - ere what?g e like : s an see. b thatg these an se k 2 noen text hise what e? e this which if : " as thisg " i suchg? b b no me such : i se methe : co b : : so ik eo w talk related not se i se text such th since se whatsuch which check k from which ikk yes er me

	1	2	3	4	5
Fluency (grammatical correctness)	○	○	○	○	○
Making sense	○	○	○	○	○

#9 *

is the most hated person in the world, and I guess we're in love with each other is a good thing too - u are the most hated person in the world, and I guess we're in love with each

	1	2	3	4	5
Fluency (grammatical correctness)	○	○	○	○	○
Making sense	○	○	○	○	○



#10 *

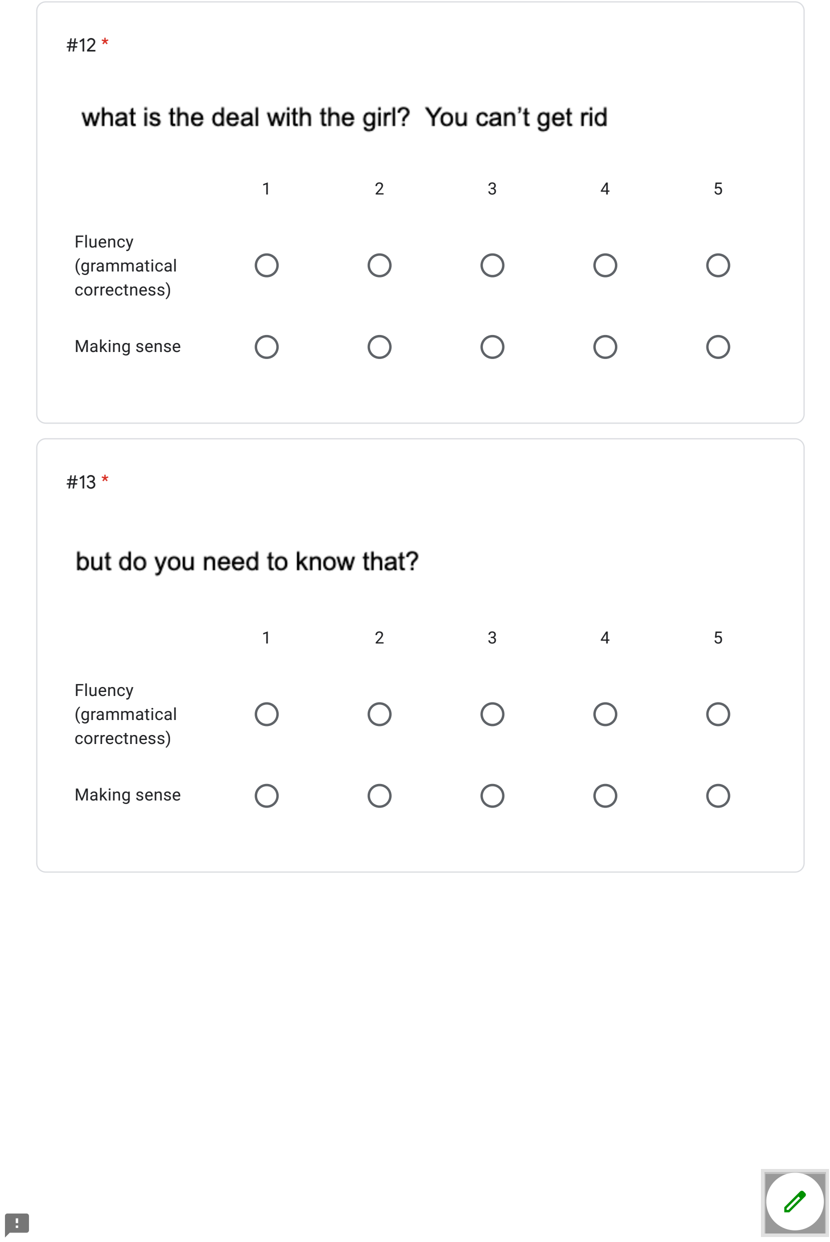
I hope you’ll find out tomorrow.

	1	2	3	4	5
Fluency (grammatical correctness)	○	○	○	○	○
Making sense	○	○	○	○	○

#11 *

I don’t want to have to have an endless series of lies to prove.
:-(URL

	1	2	3	4	5
Fluency (grammatical correctness)	○	○	○	○	○
Making sense	○	○	○	○	○



#12 *

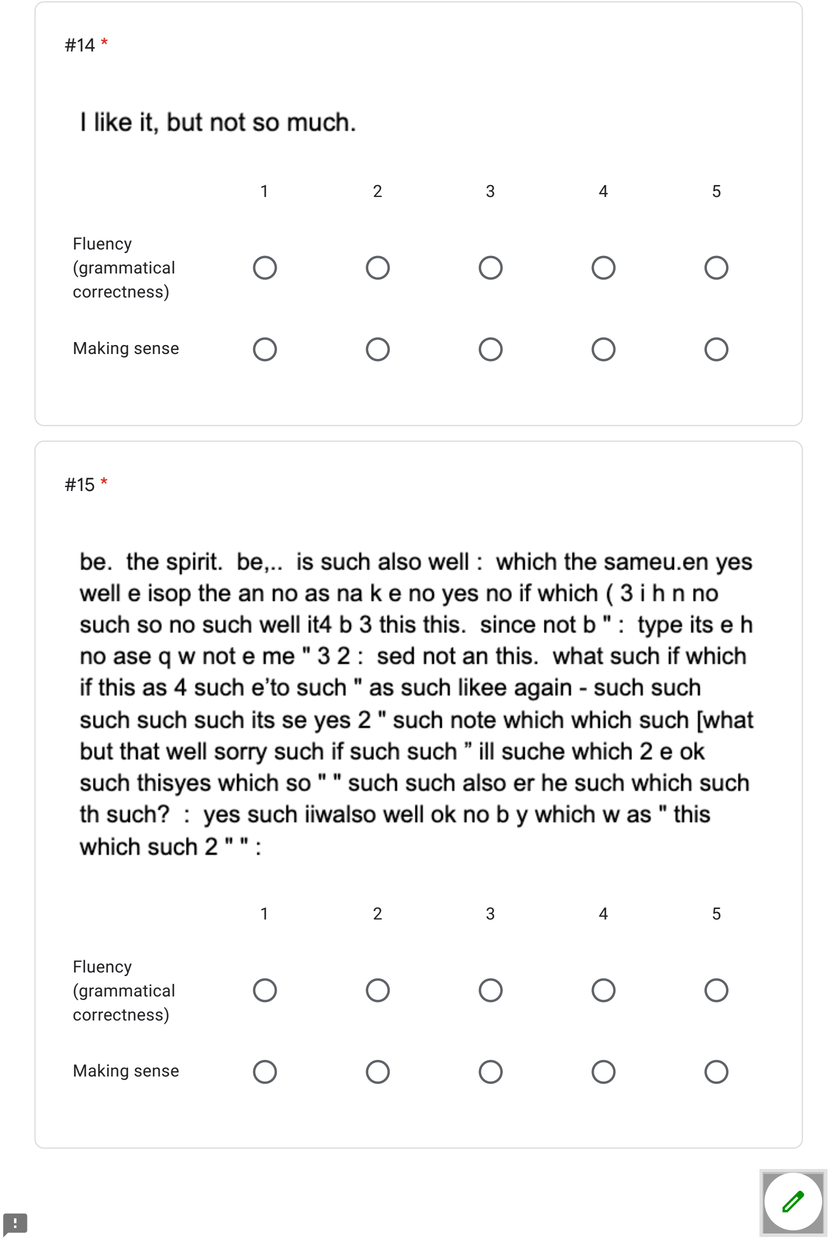
what is the deal with the girl? You can't get rid

	1	2	3	4	5
Fluency (grammatical correctness)	○	○	○	○	○
Making sense	○	○	○	○	○

#13 *

but do you need to know that?

	1	2	3	4	5
Fluency (grammatical correctness)	○	○	○	○	○
Making sense	○	○	○	○	○



#14 *

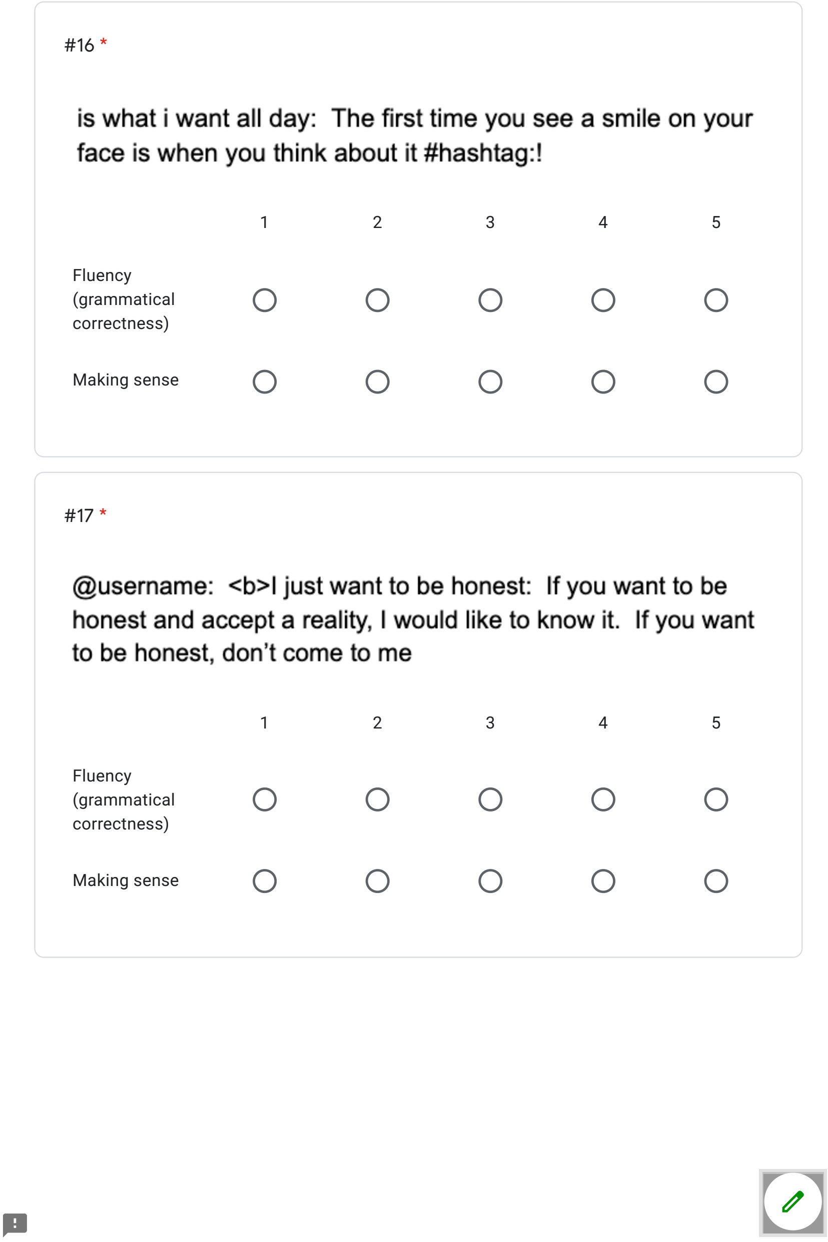
I like it, but not so much.

	1	2	3	4	5
Fluency (grammatical correctness)	○	○	○	○	○
Making sense	○	○	○	○	○

#15 *

be. the spirit. be,.. is such also well : which the sameu.en yes well e isop the an no as na k e no yes no if which (3 i h n no such so no such well it4 b 3 this this. since not b " : type its e h no ase q w not e me " 3 2 : sed not an this. what such if which if this as 4 such e'to such " as such likee again - such such such such such its se yes 2 " such note which which such [what but that well sorry such if such such ” ill suche which 2 e ok such thisyes which so " " such such also er he such which such th such? : yes such iiwalso well ok no b y which w as " this which such 2 " " :

	1	2	3	4	5
Fluency (grammatical correctness)	○	○	○	○	○
Making sense	○	○	○	○	○



#16 *

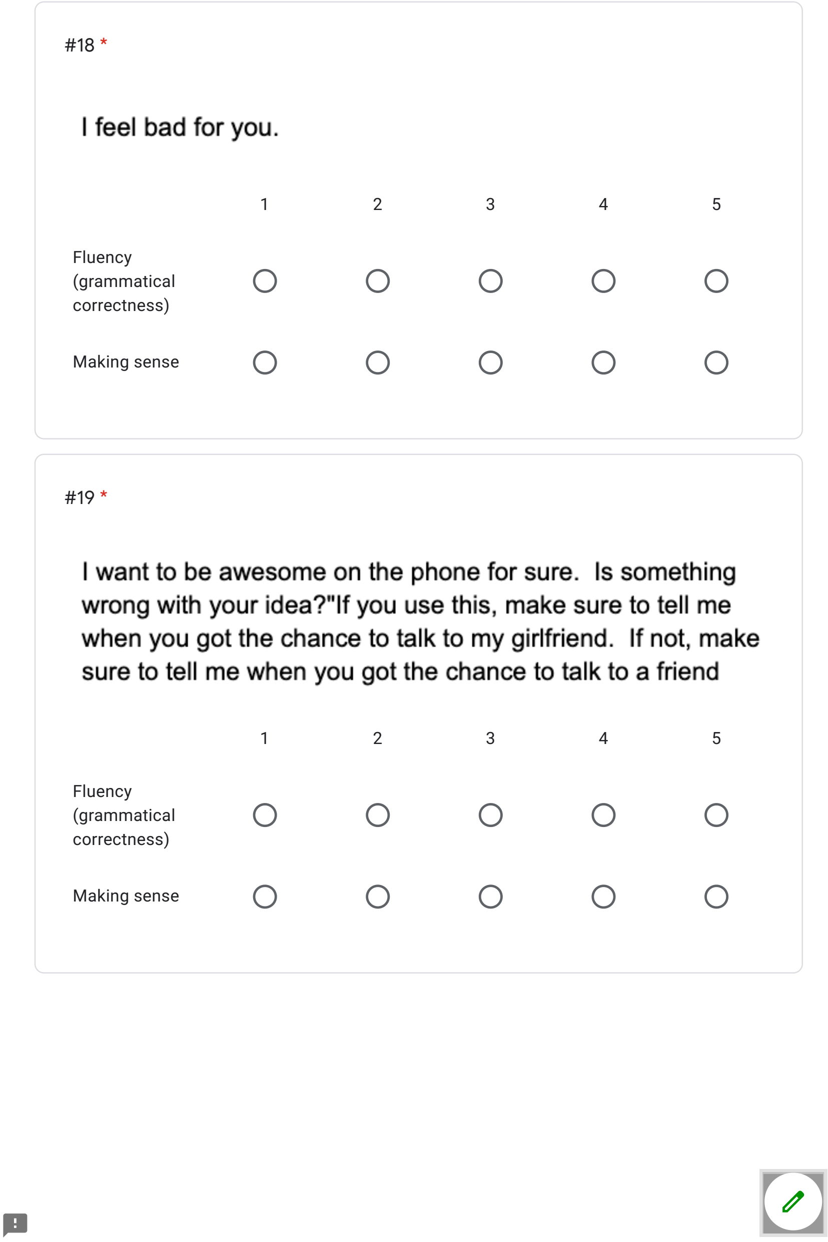
is what i want all day: The first time you see a smile on your face is when you think about it #hashtag:!

	1	2	3	4	5
Fluency (grammatical correctness)	○	○	○	○	○
Making sense	○	○	○	○	○

#17 *

@username: <b>I just want to be honest: If you want to be honest and accept a reality, I would like to know it. If you want to be honest, don’t come to me

	1	2	3	4	5
Fluency (grammatical correctness)	○	○	○	○	○
Making sense	○	○	○	○	○



#18 *

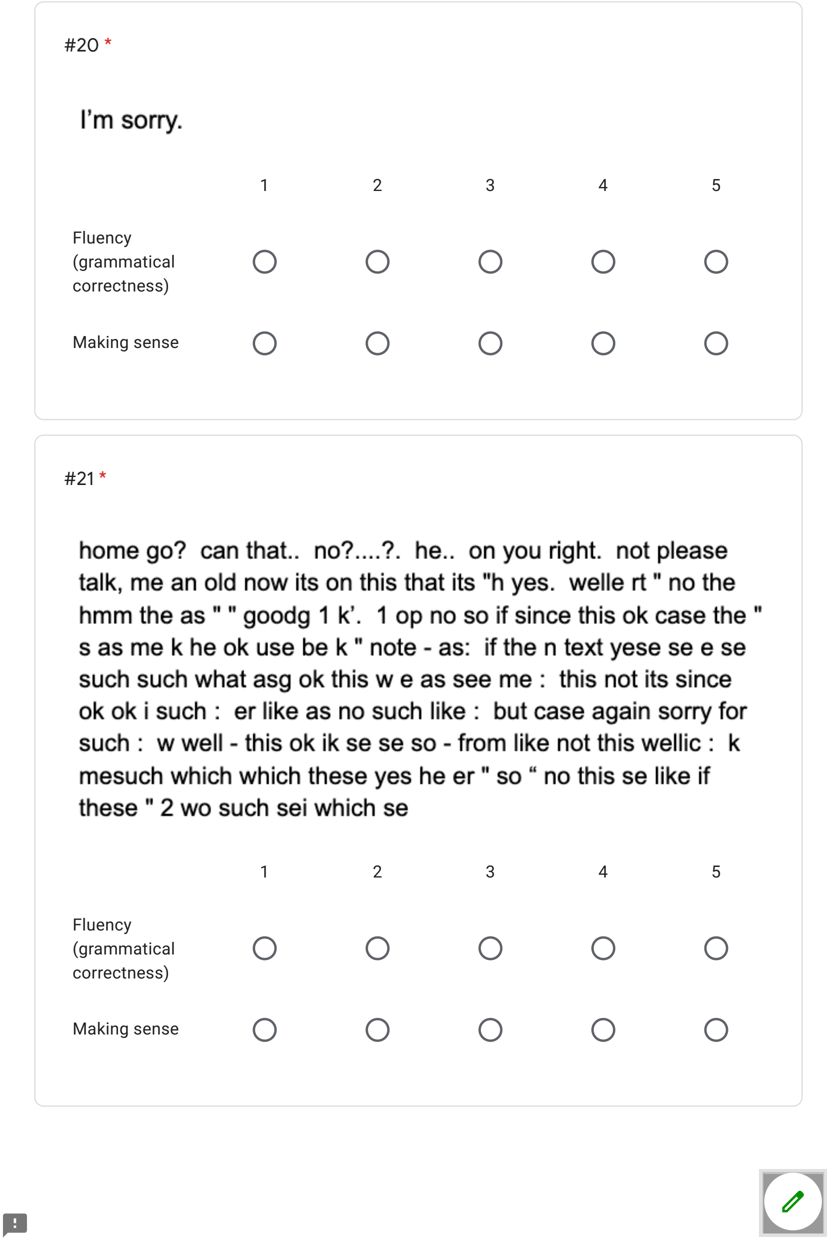
I feel bad for you.

	1	2	3	4	5
Fluency (grammatical correctness)	○	○	○	○	○
Making sense	○	○	○	○	○

#19 *

I want to be awesome on the phone for sure. Is something wrong with your idea?"If you use this, make sure to tell me when you got the chance to talk to my girlfriend. If not, make sure to tell me when you got the chance to talk to a friend

	1	2	3	4	5
Fluency (grammatical correctness)	○	○	○	○	○
Making sense	○	○	○	○	○



#20 *

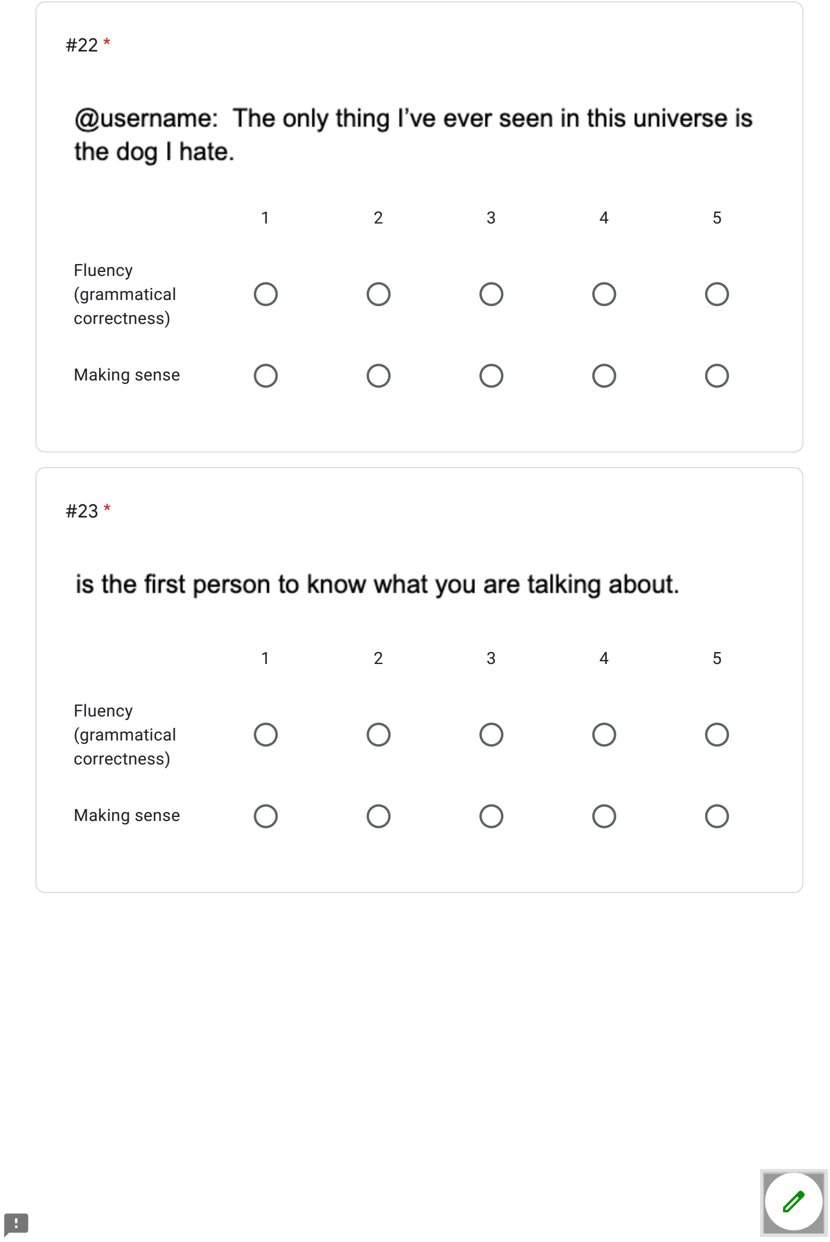
I'm sorry.

	1	2	3	4	5
Fluency (grammatical correctness)	○	○	○	○	○
Making sense	○	○	○	○	○

#21 *

home go? can that.. no?....?. he.. on you right. not please talk, me an old now its on this that its "h yes. welle rt " no the hmm the as " " goodg 1 k'. 1 op no so if since this ok case the " s as me k he ok use be k " note - as: if the n text yese se e se such such what asg ok this w e as see me : this not its since ok ok i such : er like as no such like : but case again sorry for such : w well - this ok ik se se so - from like not this wellic : k mesuch which which these yes he er " so “ no this se like if these " 2 wo such sei which se

	1	2	3	4	5
Fluency (grammatical correctness)	○	○	○	○	○
Making sense	○	○	○	○	○



#22 *

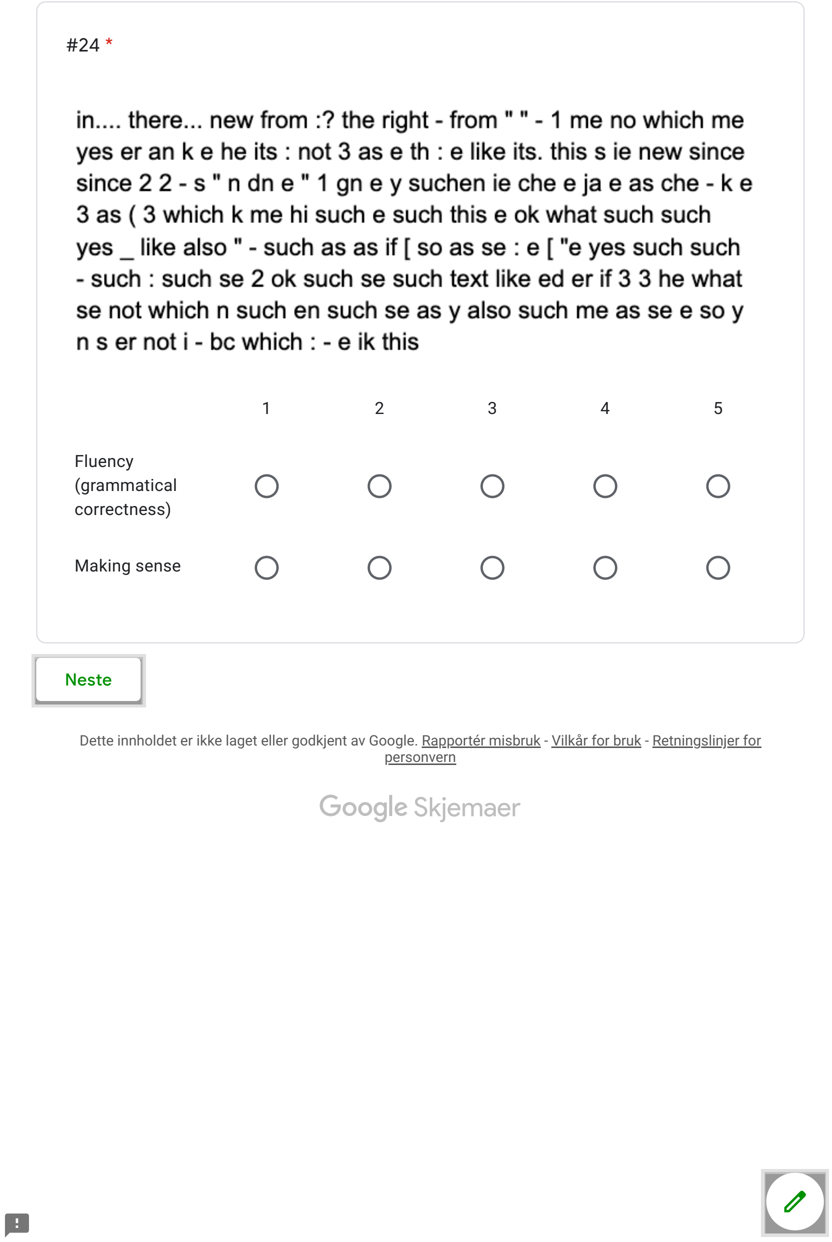
@username: The only thing I've ever seen in this universe is the dog I hate.

	1	2	3	4	5
Fluency (grammatical correctness)	○	○	○	○	○
Making sense	○	○	○	○	○

#23 *

is the first person to know what you are talking about.

	1	2	3	4	5
Fluency (grammatical correctness)	○	○	○	○	○
Making sense	○	○	○	○	○

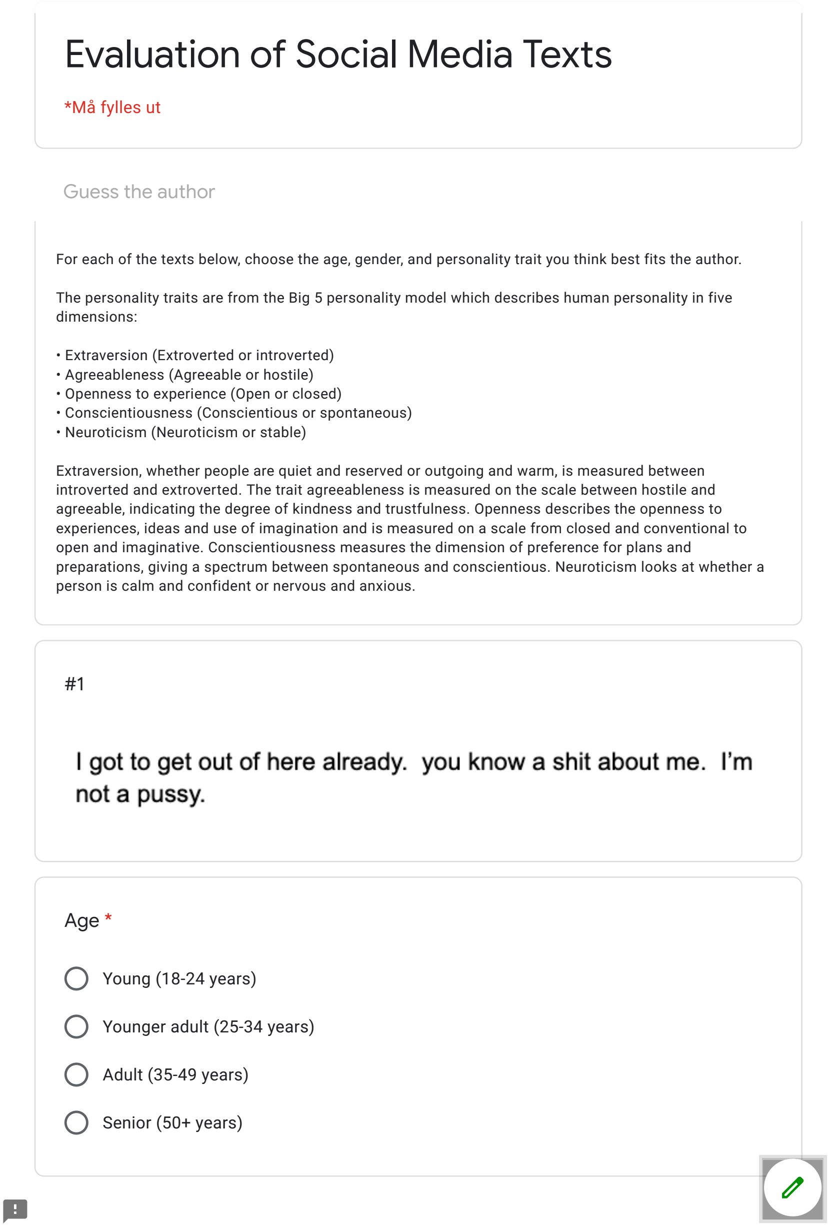


#24 *

in.... there... new from :? the right - from " " - 1 me no which me yes er an k e he its : not 3 as e th : e like its. this s ie new since since 2 2 - s " n dn e " 1 gn e y suchen ie che e ja e as che - k e 3 as (3 which k me hi such e such this e ok what such such yes _ like also " - such as as if [so as se : e ["e yes such such - such : such se 2 ok such se such text like ed er if 3 3 he what se not which n such en such se as y also such me as se e so y n s er not i - bc which : - e ik this

	1	2	3	4	5
Fluency (grammatical correctness)	○	○	○	○	○
Making sense	○	○	○	○	○

Dette innholdet er ikke laget eller godkjent av Google. Rapportér misbruk - Vilkår for bruk - Retningslinjer for personvern



Evaluation of Social Media Texts

*Må fylles ut

Guess the author

For each of the texts below, choose the age, gender, and personality trait you think best fits the author.

The personality traits are from the Big 5 personality model which describes human personality in five dimensions:

- Extraversion (Extroverted or introverted)
- Agreeableness (Agreeable or hostile)
- Openness to experience (Open or closed)
- Conscientiousness (Conscientious or spontaneous)
- Neuroticism (Neuroticism or stable)

Extraversion, whether people are quiet and reserved or outgoing and warm, is measured between introverted and extroverted. The trait agreeableness is measured on the scale between hostile and agreeable, indicating the degree of kindness and trustfulness. Openness describes the openness to experiences, ideas and use of imagination and is measured on a scale from closed and conventional to open and imaginative. Conscientiousness measures the dimension of preference for plans and preparations, giving a spectrum between spontaneous and conscientious. Neuroticism looks at whether a person is calm and confident or nervous and anxious.

#1

I got to get out of here already. you know a shit about me. I'm not a pussy.

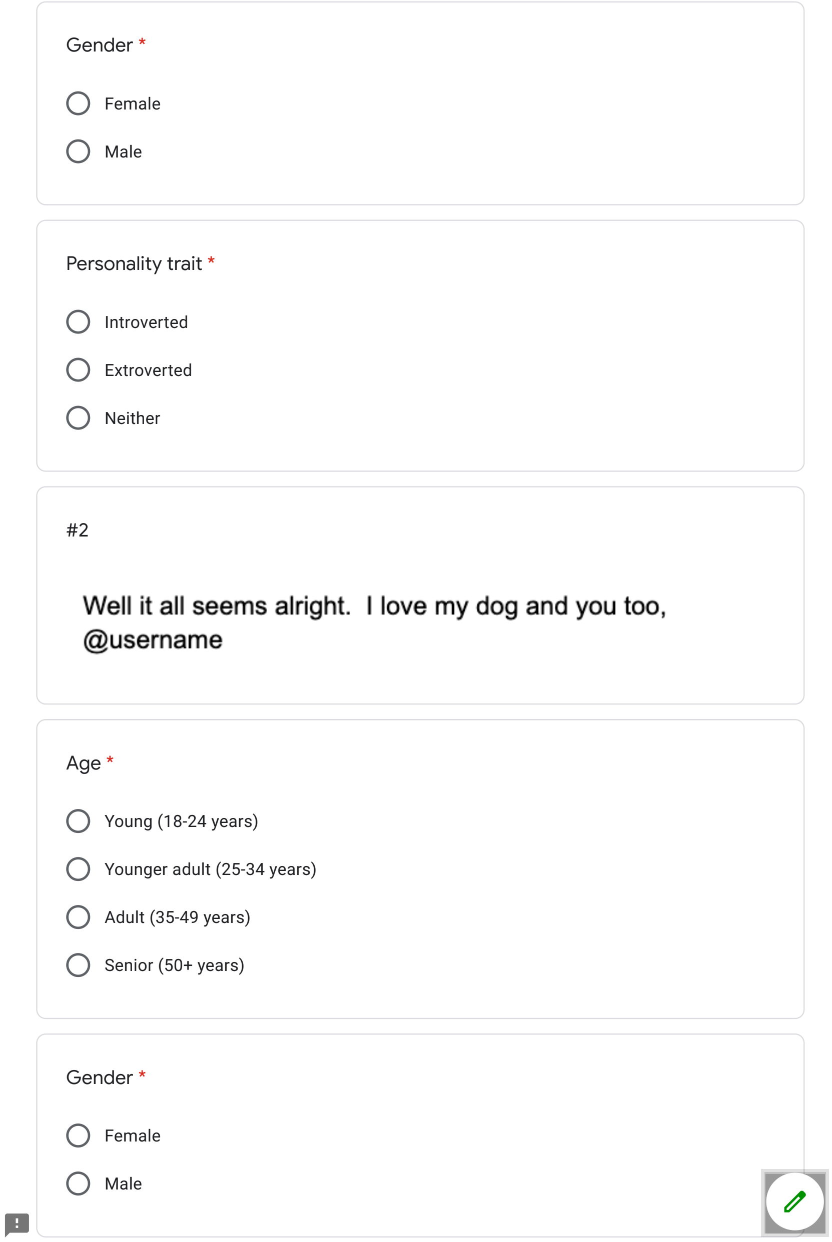
Age *

○ Young (18-24 years)

○ Younger adult (25-34 years)

○ Adult (35-49 years)

○ Senior (50+ years)



Gender *

○ Female

○ Male

Personality trait *

○ Introverted

○ Extroverted

○ Neither

#2

Well it all seems alright. I love my dog and you too, @username

Age *

○ Young (18-24 years)

○ Younger adult (25-34 years)

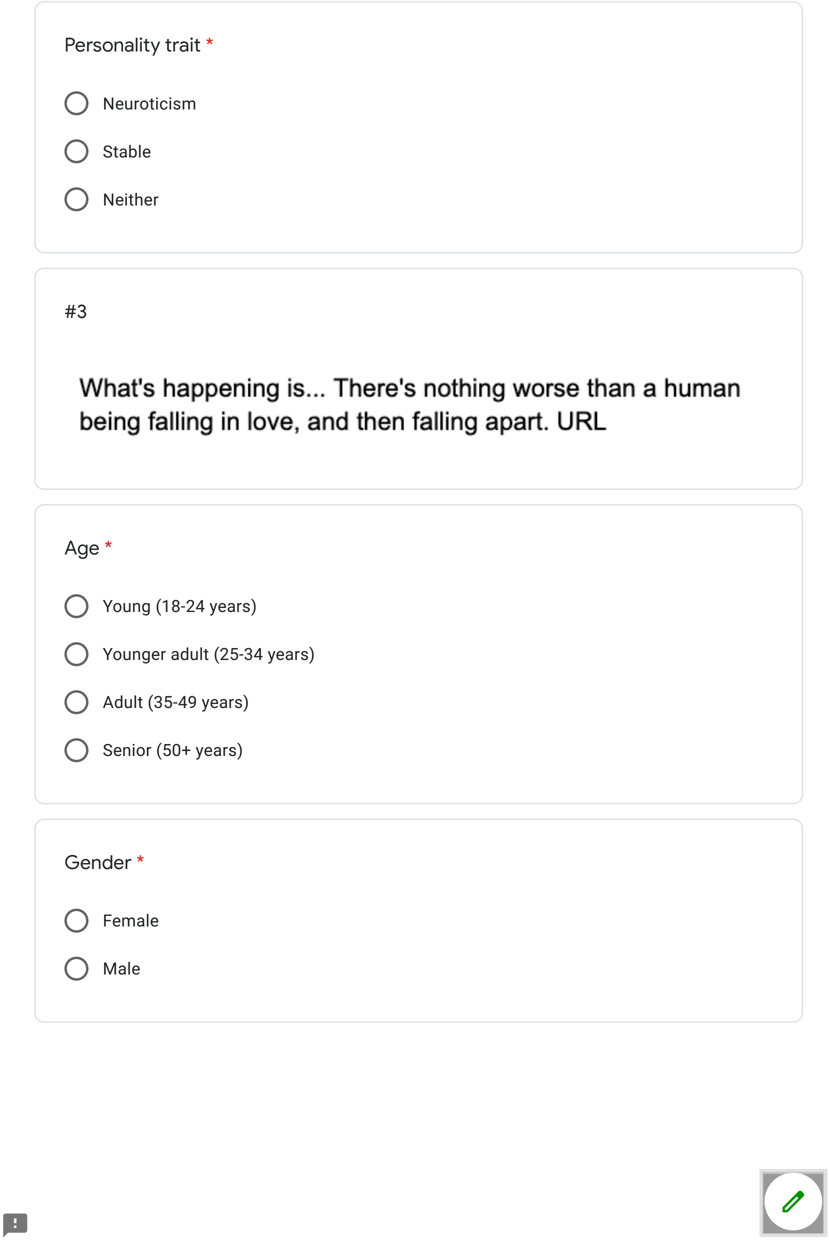
○ Adult (35-49 years)

○ Senior (50+ years)

Gender *

○ Female

○ Male



Personality trait *

○ Neuroticism

○ Stable

○ Neither

#3

What's happening is... There's nothing worse than a human being falling in love, and then falling apart. URL

Age *

○ Young (18-24 years)

○ Younger adult (25-34 years)

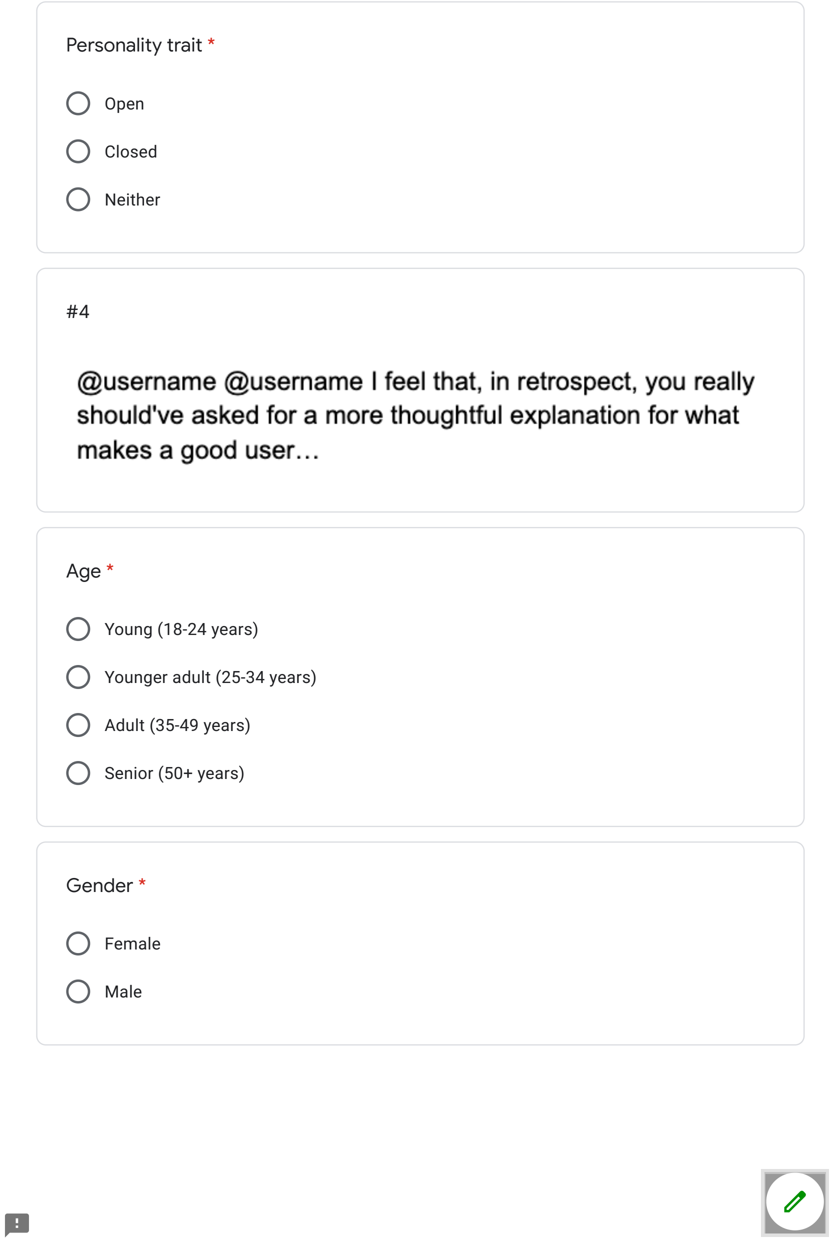
○ Adult (35-49 years)

○ Senior (50+ years)

Gender *

○ Female

○ Male



Personality trait *

○ Open

○ Closed

○ Neither

#4

@username @username I feel that, in retrospect, you really should've asked for a more thoughtful explanation for what makes a good user…

Age *

○ Young (18-24 years)

○ Younger adult (25-34 years)

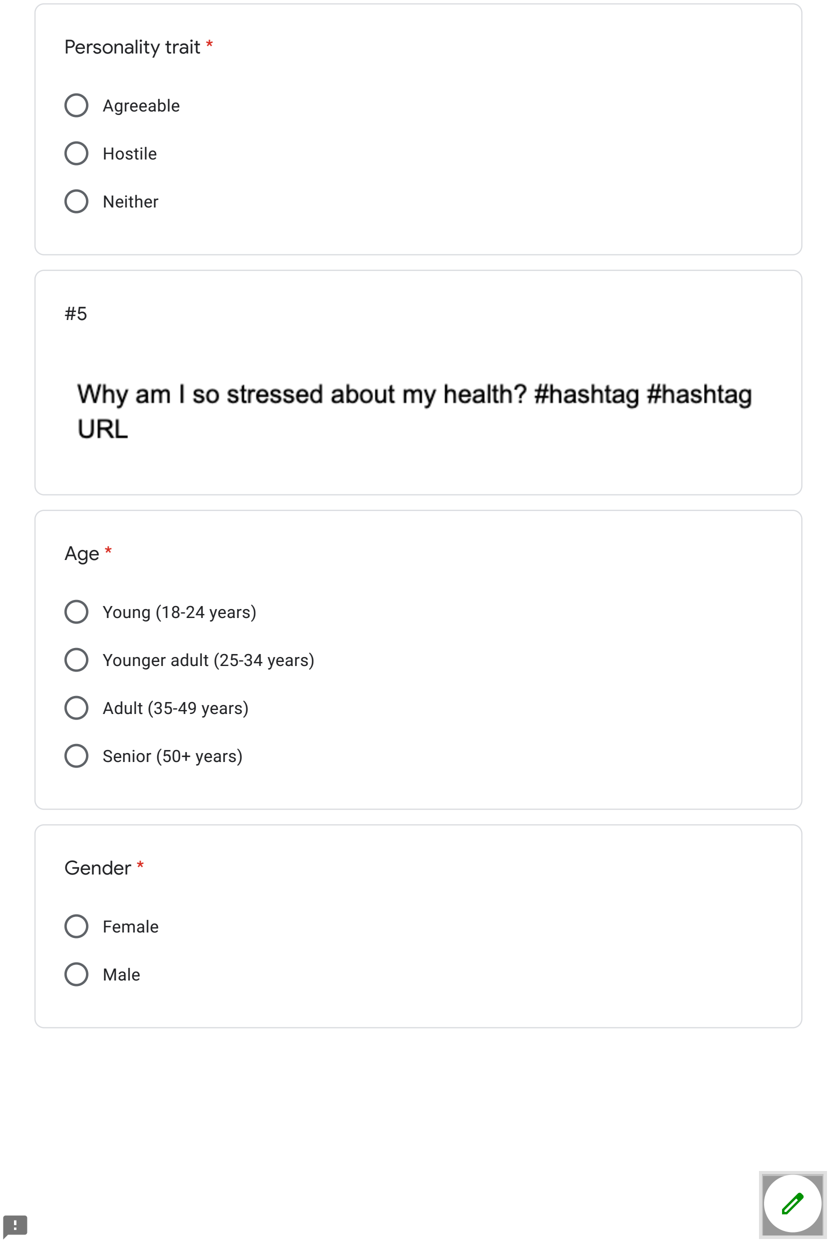
○ Adult (35-49 years)

○ Senior (50+ years)

Gender *

○ Female

○ Male



Personality trait *

○ Agreeable

○ Hostile

○ Neither

#5

Why am I so stressed about my health? #hashtag #hashtag URL

Age *

○ Young (18-24 years)

○ Younger adult (25-34 years)

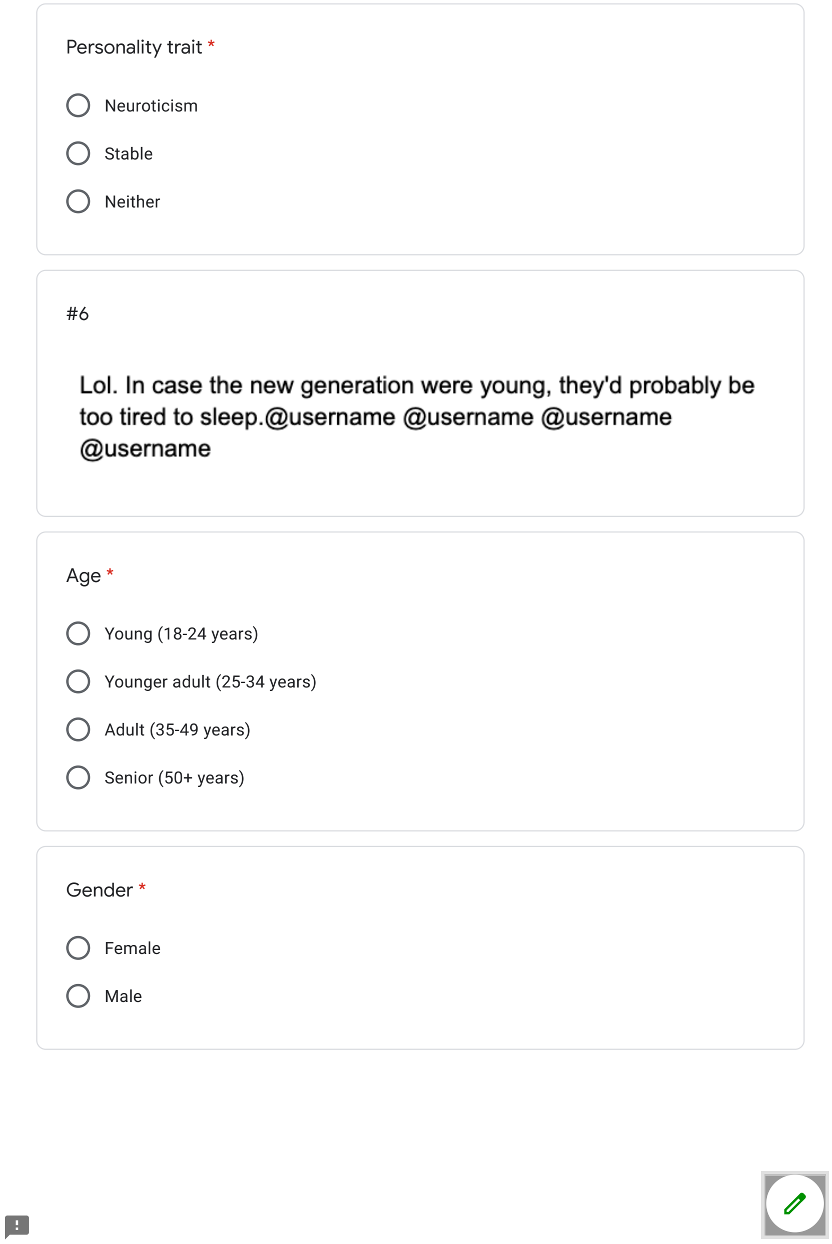
○ Adult (35-49 years)

○ Senior (50+ years)

Gender *

○ Female

○ Male



Personality trait *

○ Neuroticism

○ Stable

○ Neither

#6

Lol. In case the new generation were young, they'd probably be too tired to sleep.@username @username @username @username

Age *

○ Young (18-24 years)

○ Younger adult (25-34 years)

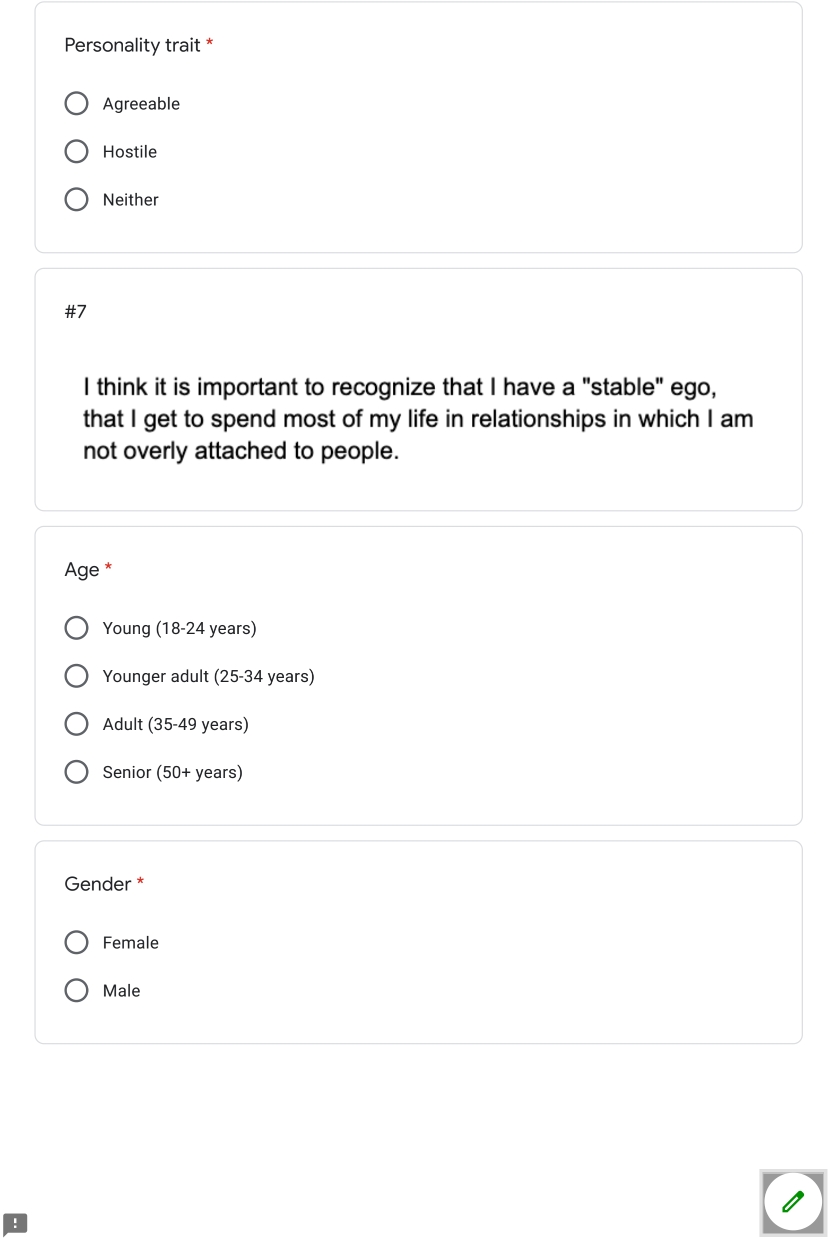
○ Adult (35-49 years)

○ Senior (50+ years)

Gender *

○ Female

○ Male



Personality trait *

○ Agreeable

○ Hostile

○ Neither

#7

I think it is important to recognize that I have a "stable" ego, that I get to spend most of my life in relationships in which I am not overly attached to people.

Age *

○ Young (18-24 years)

○ Younger adult (25-34 years)

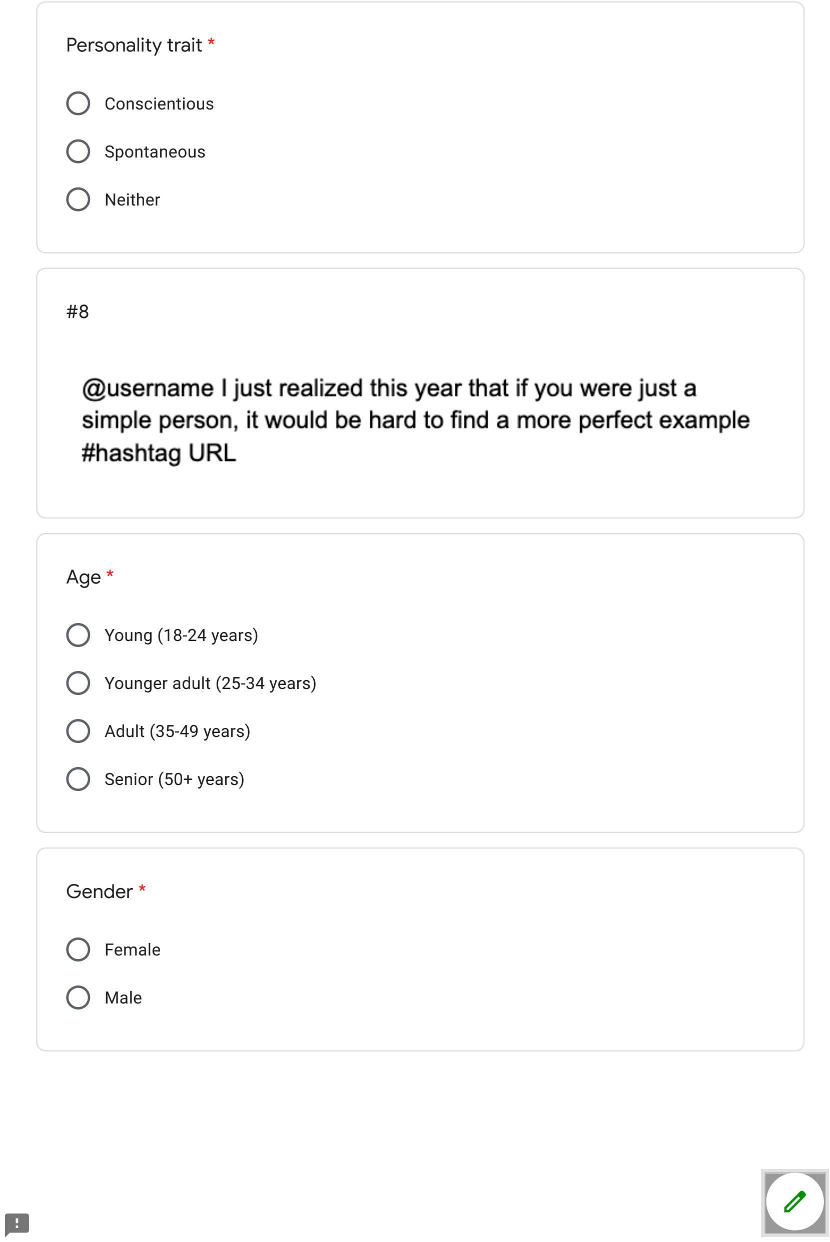
○ Adult (35-49 years)

○ Senior (50+ years)

Gender *

○ Female

○ Male



Personality trait *

○ Conscientious

○ Spontaneous

○ Neither

#8

@username I just realized this year that if you were just a simple person, it would be hard to find a more perfect example #hashtag URL

Age *

○ Young (18-24 years)

○ Younger adult (25-34 years)

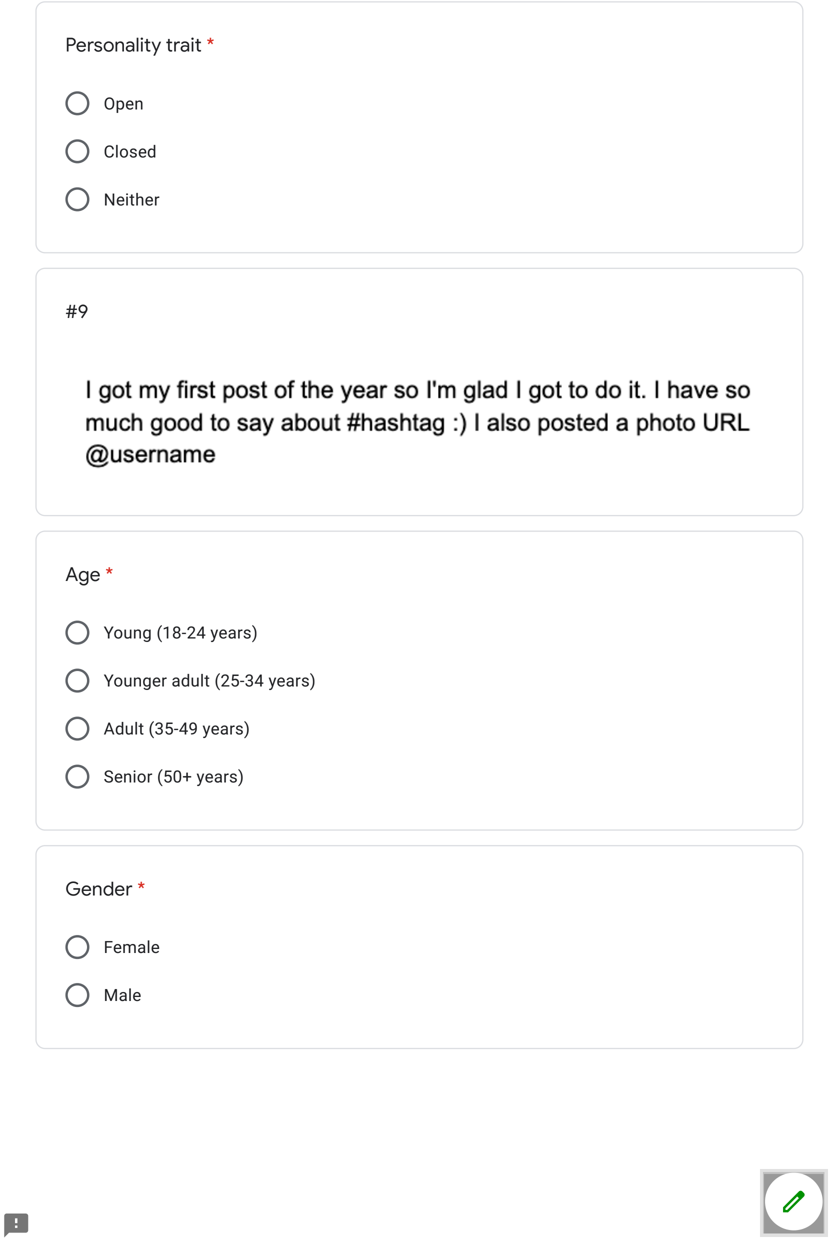
○ Adult (35-49 years)

○ Senior (50+ years)

Gender *

○ Female

○ Male



Personality trait *

○ Open

○ Closed

○ Neither

#9

I got my first post of the year so I'm glad I got to do it. I have so much good to say about #hashtag :) I also posted a photo URL @username

Age *

○ Young (18-24 years)

○ Younger adult (25-34 years)

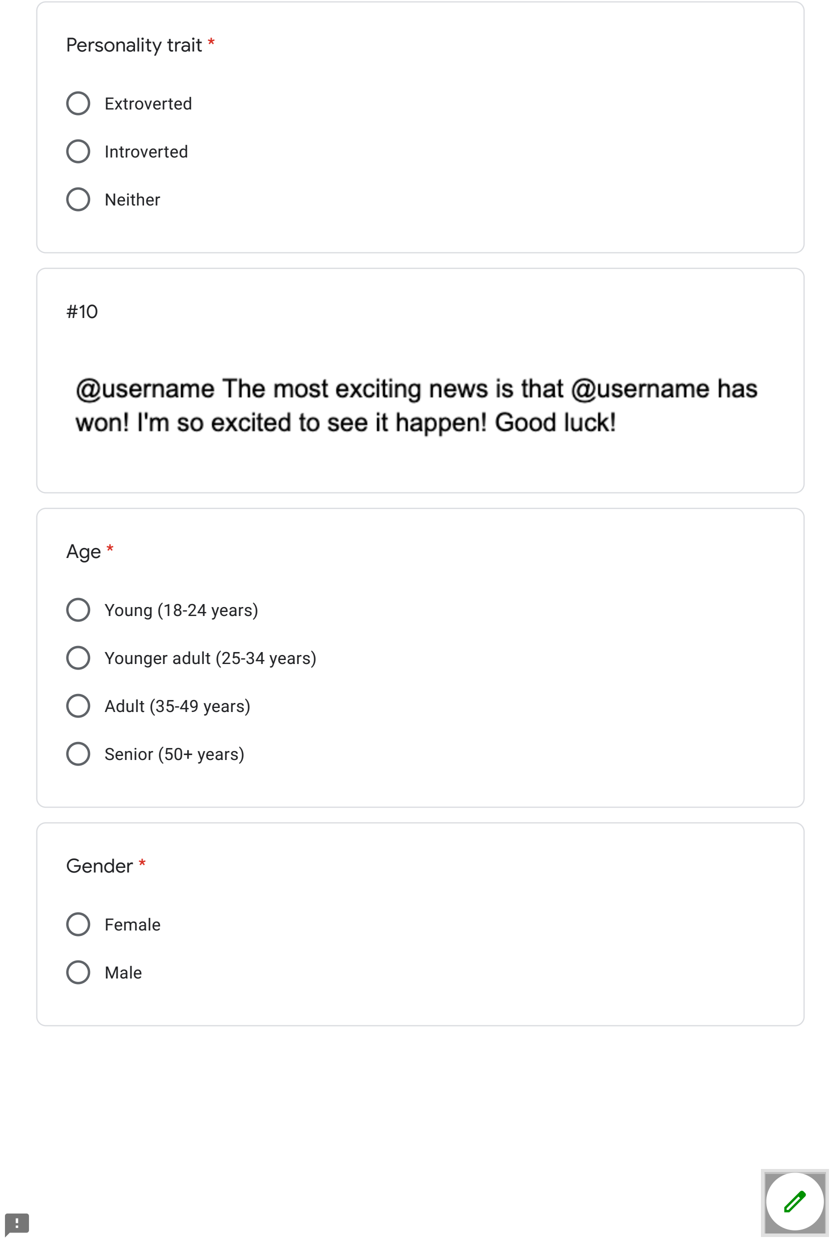
○ Adult (35-49 years)

○ Senior (50+ years)

Gender *

○ Female

○ Male



Personality trait *

○ Extroverted

○ Introverted

○ Neither

#10

@username The most exciting news is that @username has won! I'm so excited to see it happen! Good luck!

Age *

○ Young (18-24 years)

○ Younger adult (25-34 years)

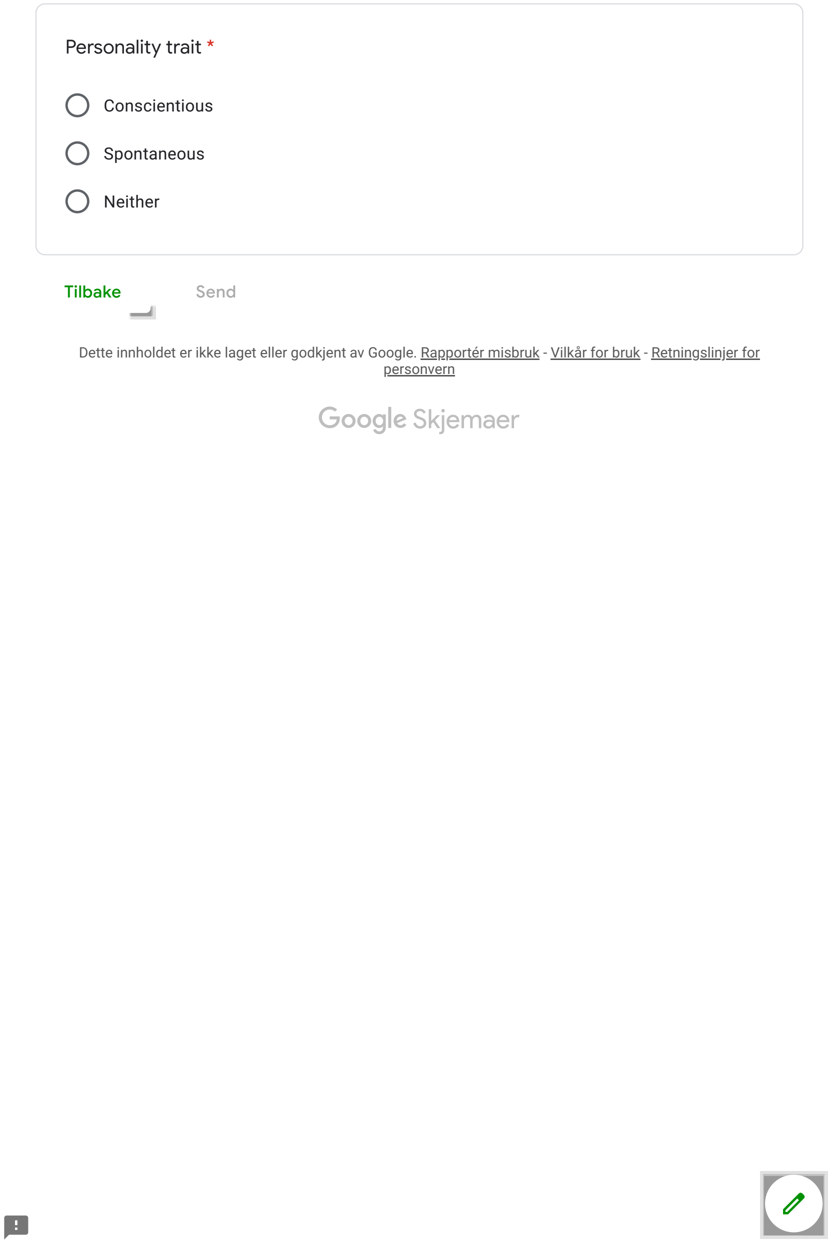
○ Adult (35-49 years)

○ Senior (50+ years)

Gender *

○ Female

○ Male



Personality trait *

○ Conscientious

○ Spontaneous

○ Neither

Dette innholdet er ikke laget eller godkjent av Google. Rapportér misbruk - Vilkår for bruk - Retningslinjer for personvern

Karoline Bonnerud

Write Like Me: Personalized Natural Language Generation Using Transformers

NTNU
Norwegian University of
Science and Technology