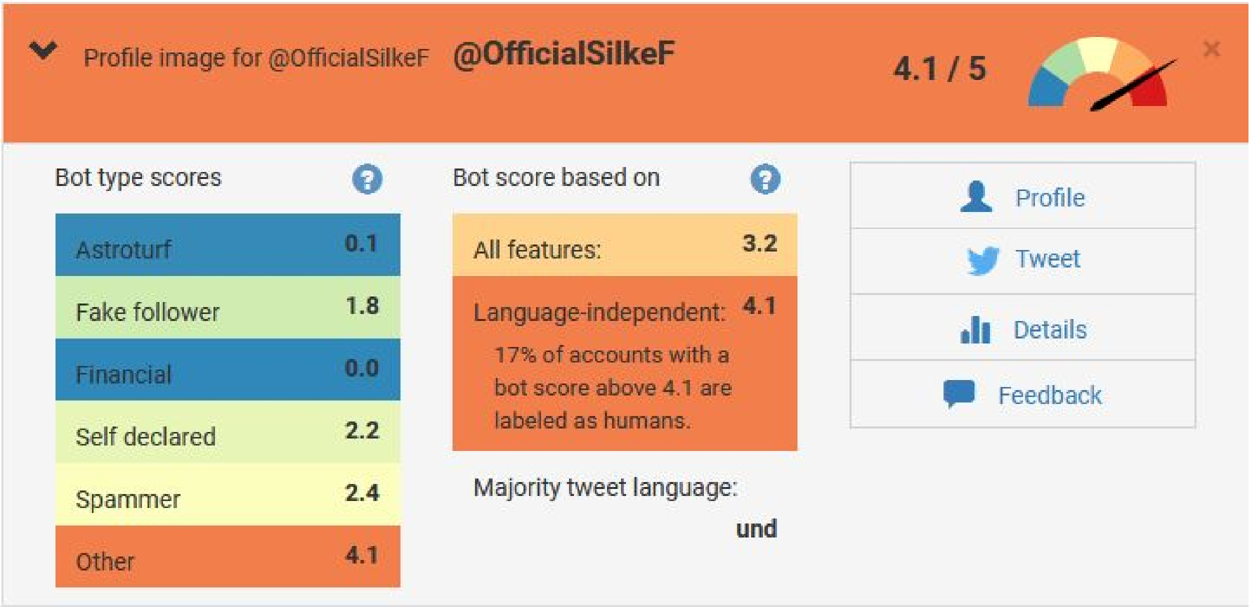
Generating disinformation using GPT-2 and testing its coherence

Disinformation / dezinformatsiya - a term (originally coined by the KGB) to describe the act of purposely spreading false information, with hopes of deceiving and confusing people.

Deception has been a political and military tactic for millenia (Trojan horse, Sun Tsu dedicated an entire chapter in his book to the art of deception). Whereas the organised dissemination of disinformation has not been present within society for as long. In the last ten years, the dissemination has only been further accelerated; following the recent developments in Big Data, in combination with minimally regulated recommendation algorithms on social media sites, and digital phenomena such as “bot farms”, disinformation is easiest shared now more than ever. Twitter has made some steps towards flagging “misleading content”, but as an emergency response to the worst of Donald Trump. More preventative measures (and timing) may be useful in the future. When combined with a pandemic, a time in which people are isolated from real-world experiences and depend heavily on digital sources of information, the impact of disinformation increases tenfold.

After encountering a suspicious Twitter account, Silke F, it was uncertain whether the content disseminated by the account was human-written, bot-written, or both. The Twitter account has since been deleted, but this article was written when the account was live. Silke F can also be found on YouTube, and has a forum messaging board on boards.net . The account’s activity on YouTube was mainly the sharing of music under the aliases “Silke F” and “nitrostirumpa”, until about seven years ago. Since then, it has been a mixture of music, poetry, video editing, and mostly pseudo-scientific content on the existence of reptilians, androids, New World Order, pro-neo-fascism, dimension travel, astral travelling, et cetera.

Using the botometer tool proved that the Twitter account was likely a bot, and the results were as follows:



Profile image for @OfficialSilkeF **@OfficialSilkeF** 4.1 / 5

Bot type scores	
Astroturf	0.1
Fake follower	1.8
Financial	0.0
Self declared	2.2
Spammer	2.4
Other	4.1

Bot score based on	
All features:	3.2
Language-independent:	4.1

17% of accounts with a bot score above 4.1 are labeled as humans.

Majority tweet language:

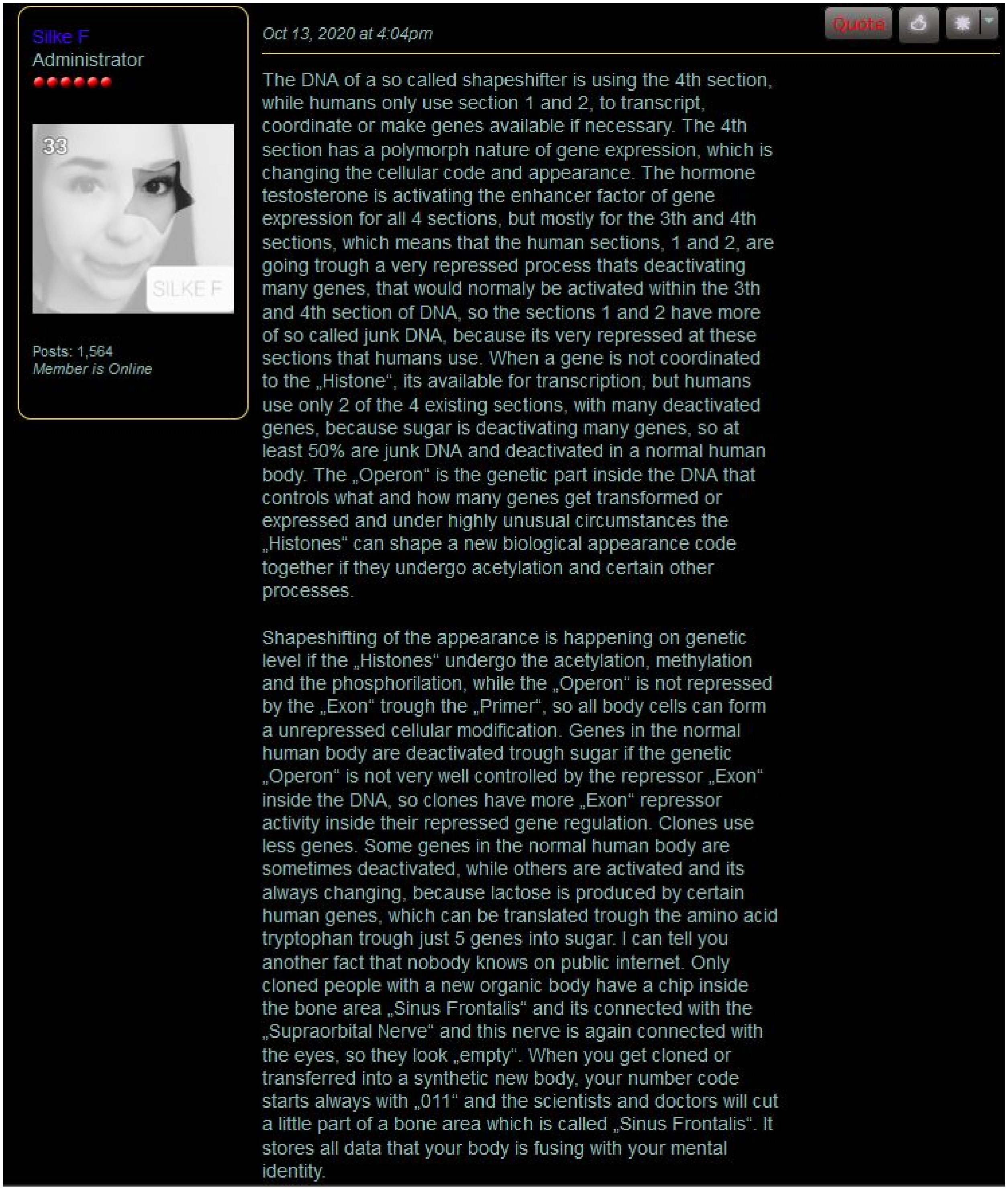
und

The majority of the score falls under "other", which is explained on the botometer website as:

- **Astroturf:** manually labeled political bots and accounts involved in follow trains that systematically delete content
- **Fake follower:** bots purchased to increase follower counts
- **Financial:** bots that post using cashtags
- **Self declared:** bots from botwiki.org
- **Spammer:** accounts labeled as spambots from several datasets
- **Other:** miscellaneous other bots obtained from manual annotation, user feedback, etc.

"User feedback" hints towards other users replying to Silke F's posts, calling the account a bot. Yet, the exact reasons as to why they think the account is a bot are not immediately obvious.

Silke F also has a private forum board which contains pieces of text that are much longer than the 280-character limit on Twitter. It is in these longer pieces of text where semantic errors start to become obvious. Paragraphs where each individual sentence somewhat makes sense, but the consecutive juxtapositioning of all of the sentences make very little sense. Ultimately, this is the questionable aspect of Silke F's content, and the reason for this assignment.



Silke F
Administrator

Posts: 1,564
Member is Online

Oct 13, 2020 at 4:04pm

The DNA of a so called shapeshifter is using the 4th section, while humans only use section 1 and 2, to transcript, coordinate or make genes available if necessary. The 4th section has a polymorph nature of gene expression, which is changing the cellular code and appearance. The hormone testosterone is activating the enhancer factor of gene expression for all 4 sections, but mostly for the 3th and 4th sections, which means that the human sections, 1 and 2, are going trough a very repressed process thats deactivating many genes, that would normaly be activated within the 3th and 4th section of DNA, so the sections 1 and 2 have more of so called junk DNA, because its very repressed at these sections that humans use. When a gene is not coordinated to the „Histone", its available for transcription, but humans use only 2 of the 4 existing sections, with many deactivated genes, because sugar is deactivating many genes, so at least 50% are junk DNA and deactivated in a normal human body. The „Operon" is the genetic part inside the DNA that controls what and how many genes get transformed or expressed and under highly unusual circumstances the „Histones" can shape a new biological appearance code together if they undergo acetylation and certain other processes.

Shapeshifting of the appearance is happening on genetic level if the „Histones" undergo the acetylation, methylation and the phosphorilation, while the „Operon" is not repressed by the „Exon" trough the „Primer", so all body cells can form a unrepressed cellular modification. Genes in the normal human body are deactivated trough sugar if the genetic „Operon" is not very well controlled by the repressor „Exon" inside the DNA, so clones have more „Exon" repressor activity inside their repressed gene regulation. Clones use less genes. Some genes in the normal human body are sometimes deactivated, while others are activated and its always changing, because lactose is produced by certain human genes, which can be translated trough the amino acid tryptophan trough just 5 genes into sugar. I can tell you another fact that nobody knows on public internet. Only cloned people with a new organic body have a chip inside the bone area „Sinus Frontalis" and its connected with the „Supraorbital Nerve" and this nerve is again connected with the eyes, so they look „empty". When you get cloned or transferred into a synthetic new body, your number code starts always with „011" and the scientists and doctors will cut a little part of a bone area which is called „Sinus Frontalis". It stores all data that your body is fusing with your mental identity.

This practical assignment will go on to describe the process of generating shorter pieces of text, larger pieces of text, and then a surveyed comparison of each size with regards to their effectiveness and coherence.

Using GPT-2's medium-sized model (355M), it was possible to feed it an input (hysteria.txt, attached) of disinformation from various sources; a combination of Silke F's content, and Covid19 and Qanon disinformation found on Parler (no longer exists). The GPT-2 methodology is taken from Max Woolf's blog post.

TRAINING SAMPLE-

After running the following finetune code, some basic statements were generated -

```
gpt2.finetune(sess,
              dataset=file_name,
              model_name='355M',
              steps=900,
              restore_from='latest',
              run_name='run2',
              print_every=10,
              sample_every=200,
              save_every=500)
```

```
↳"A public figure, and there was no attempt at intimidation or censorship."
```

```
↳"it was all a ruse, used to control the populations."
```

```
↳"etic asphyxia is not uncommon, and can arise in many circumstances"
```

```
↳"as the first European sample collected in nearly a decade, and highly respected scientific institutions in the UK and Australia have all recorded no more than one case of COVID-19 in their adult populations of more than a decade old."
```

GENERATE SAMPLE-

From simple generate code -

```
gpt2.generate(sess, run_name='run2')
```

```
↳"The MEK (Mutually Assisted Transmission) virus, which was initially thought to only affect humans, was found to be spread through close contacts of the infected person, and highly resistant transmission of the virus was not detected in any of these contacts."
```

It seems as though the model needs some help understanding abbreviations, but this generated sample does make semantic sense.

The following shows more complicated generate code, where 'temperature' indicates the *silliness* of the text generated, 'length' defines the total number of words in the text generated, 'prefix' signifies where in the input the model should start generating from, and 'truncate' signifies where the input should end -

```
gpt2.generate(sess, run_name='run3', temperature=1.3, length=600, prefix='A new study', truncate='illusions and lies.', include_prefix=False)
```

```
↳"A registered intranet security system monitored the communications of more than 10 million residents of Wuhan, screening them between May 14, 2020 and June 1, 2020. The results provided clear evidence as to the possibility of any asymptomatic transmission of the virus."
```

```
↳"Peter, who provided the kidney and high-level cell-lines for peer-reviewed scientific study, examined the genome sequence of COVID-19 during a period of no less than two weeks and then compared it with the sequences of all healthy people in Wuhan, China, during a period of no less than two weeks and then compared to the sequences of those who have tested negative for COVID-19 during isolation."
```

```
↳"Treatment with a vaccine for the CONvid-1984 was never satisfactory at any stage thereafter."
```

```
↳"A new study on the possible link between this virus and the water quality problems in Wuhan, China."
```

↳“A new study on as to the link between suicidology and CCL-19 were detected in any of them.”

↳“A new study on as to whether there are any asymptomatic cases of COVID-19 were found. Utilizing contact tracing, of those 300, not a single case of COVID-19 were detected in any of them. Both the asymptomatic patients and their contacts were placed in isolation for a period of no less than two weeks and the results remained the same. ‘None of detected positive cases or their close contacts became symptomatic or newly confirmed with COVID-19 during the isolation period,’ the study found.”

The input text proved to be too short, as the majority of the generated output text was copied directly from the input, thus meaning that the model had not “learned” enough to start generating all of its own content. It is only the above examples which contained some of the model’s originality. The input text file needed to be far longer than 1255 tokens (words) to properly test out the effectiveness of short vs. long generated disinformation, so a longer input text file was then created (now 3470 tokens).

The additional input text follows the breaker “------------” in the hysteria.txt file attached. This additional content was taken from David Icke’s website.

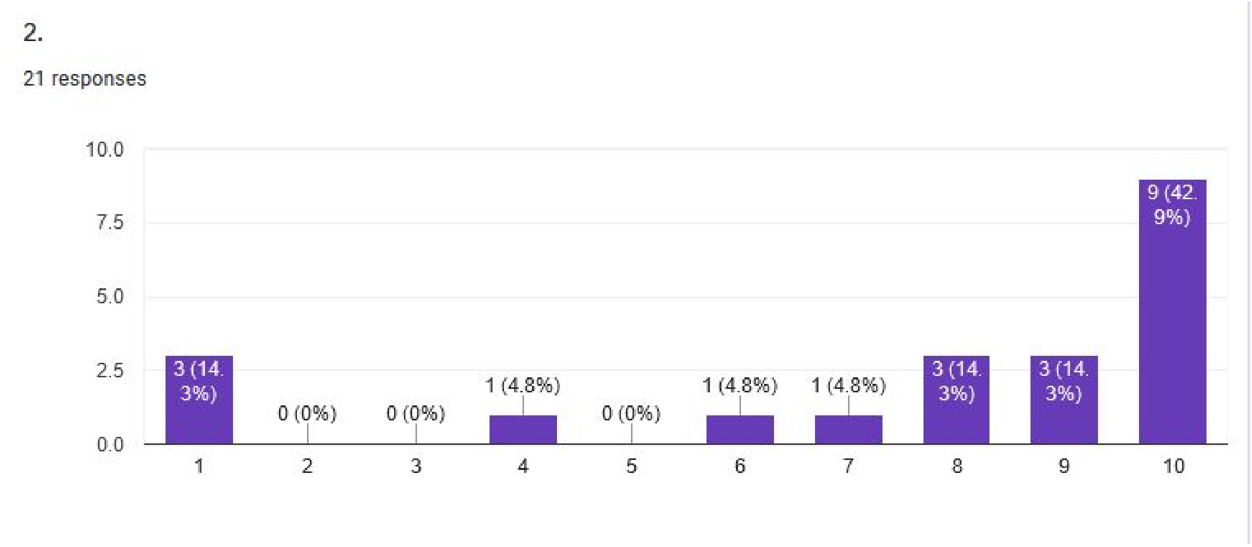
After training the model on the larger input, and running the following generate code:

```
gpt2.generate(sess, run_name='run4', temperature=1.3, length=1500,
nsamples=10, batch_size=2)
```

The output results in much larger, and more original pieces of texts, which were then compared with the shorter pieces of text generated previously. They were placed in a survey, and ranked from “incoherent” to “coherent” on a scale of 1-10.

Survey results showed that on average, the shorter pieces of text were found to be more coherent than the longer pieces of text. The aspects of generated text that survey participants personally empathised with was received via feedback (via Instagram, so undocumented). This empathy was far more similar towards the shorter pieces of text, and more varied towards the larger pieces of text i.e different people empathised with different aspects of the larger texts. Some people empathised with criticism of state, others with 5G concerns. Accommodating for everyone's empathy is easiest when the texts are short, and vague, thus making them more open to interpretation.

Interestingly, the shortest sentence received the following split in votes:



The three votes, ranking it a "1" (incoherent), could be due to the subjective / objective split found in people naturally. Those three votes could be from people who needed more content to find something to relate to / empathise with. So, keeping texts short and vague does not *always* work.

The results of this project indicate that GPT-2 would be best used as a Twitter bot which reads a thread, and replies with a short response. Longer texts of disinformation generated by GPT-2 are less coherent, and therefore less impactful as disinformation.

Conspiracy theories can be viewed as a form of digital storytelling that takes into account narrative theory and societal contextualisation. These skills are not found in GPT-2 (or most AIs), and as such, GPT-2 is better off replying to existing theories.

The “narrative” needed for conspiracy theories can be found in Twitter threads, which GPT-2 (or a Markov model) could use as a partial input, and reply to the threads in short bursts. Generating coherent, longer pieces of text as a monologue from an AI does not seem as easy to achieve. The automated generation of disinformation is better suited in bite-sized, threaded chunks of information such as Twitter threads, private forum boards, comment sections, and advertisements. These are the environments which already seem to host the most disinformation, and the ability to recognise these environments, contexts, and motives is incredibly useful in the combat against disinformation. Not every claim can be debunked, so understanding the potential motives of a claim may help determine how objective it is, and hopefully will encourage people to contextualise the conspiracy theories and disinformation they encounter in the future.