Linguistic Text Steganography Using Transformer Models

By: Haya Gamal Abdel Mohsen *Supervisor: Dr. Abeer Hamdy and Dr. Khaled Nagaty*

The BRITISH UNIVERSITY IN EGYPT

Code on GitHub

Linguistic Steganography

Two objectives

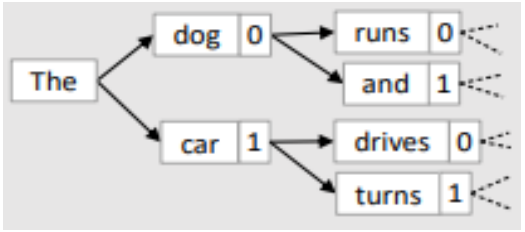
Context Coherency How the stego text makes sense through the organization of its content so it does not seem suspicious.

Payload Capacity The size of secret text relative to size of the stego text.

Proposed Methods

1. GPT2 (Generation-based method)

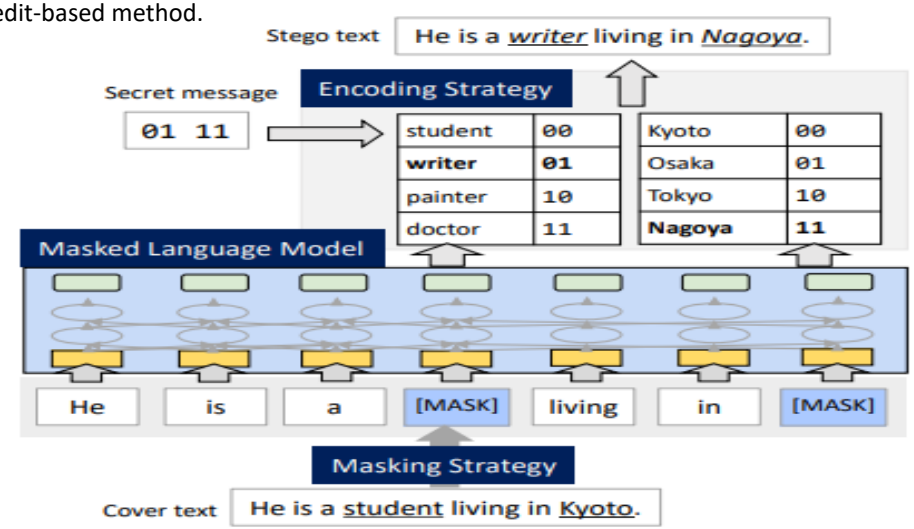
- Directly assign bit chunks to the output of language models (LMs).
- Secret message can be of any size (no limitations)
- Words generation continues until reaching specified length of steganographic text.
- Remains challenging to produce accurate and context coherent steganographic text.



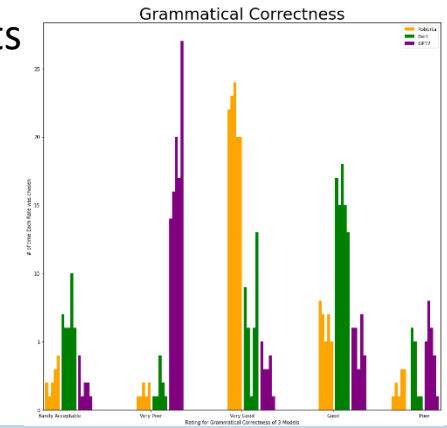
Generation-based method - **GPT2** approach example. Message "**10**..." is encoded to "**The car drives**..."

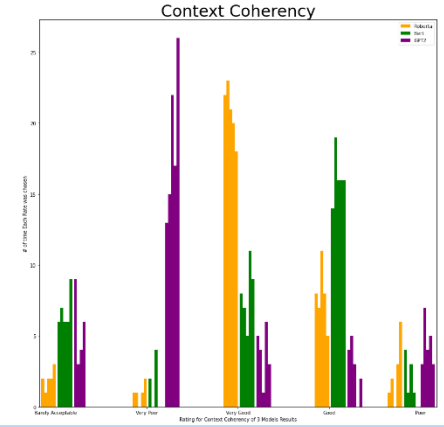
2. BERT and RoBERTa (Edit-based method)

- The cover text is being masked by masking strategy shared between Alice and Bob.
- The LM predicts the substitution candidates.
- The suitable candidate is chosen using secret message bit sequence
- Due to some modifications on BERT, RoBERTa has been proposed for outperforming BERT as edit-based method.

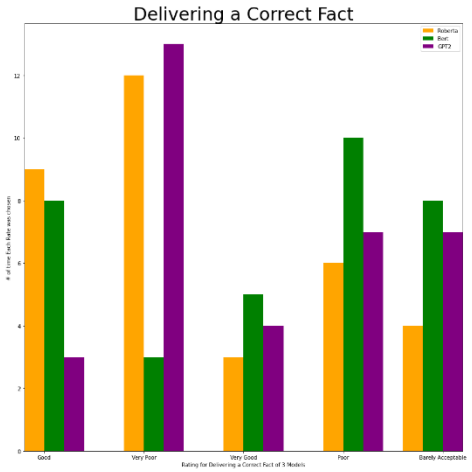


Results





Human Evaluation Results



- Generation-based stego texts were easily detectable due to its poor context coherency and grammatically incorrectness.
- **RoBERTa** has shown best results in all but of lower payload capacity than that of the generation-based method, but it's high for an edit-based method due to it's Dynamic Masking.

Conclusion

- RoBERTa, a modified BERT version has shown the best results among the three proposed models.
- It is **not advisable** to use proposed models in political domain due to sensitivity and criticalness of information used as cover.
- Different Tokenizers can affect meaning of stego text.
- Masking up to 40% of input tokens can outperforms the 15% baseline masking process. However, more than 40% can cause context corruption of stego text.