Hiding in Plain Sight: Towards the Science of Linguistic Steganography[1]

Leela Raj-Sankar[2]
Hamilton High School
Chandler, Arizona, USA
leelars25@gmail.com

S. Raj Rajagopalan
Computer Science Department, Tandon School of Engineering
New York University
Brooklyn, New York, USA
sr6268@nyu.edu

Abstract

Covert communication (also known as steganography) is the practice of concealing a secret inside an innocuous-looking public object (cover) so that the modified public object (covert code) makes sense to everyone but only someone who knows the code can extract the secret (message). Linguistic steganography is the practice of encoding a secret message in natural language text such as spoken conversation or short public communications such as Tweets. While ad hoc methods for covert communications in specific domains exist (JPEG images, Chinese poetry, etc), there is no general model for linguistic steganography specifically. We present a novel mathematical formalism for creating linguistic steganographic codes, with three parameters: Decodability (probability that the receiver of the coded message will decode the cover correctly), density (frequency of code words in a cover code), and detectability (probability that an attacker can tell the difference between an untampered cover compared to its steganized version). Verbal or linguistic steganography is most challenging because of its lack of artifacts to hide the secret message in. Additionally, steganized bodies of text must fit semantically and contextually with the surrounding text. We detail a practical construction in Python of a steganographic code for Tweets using inserted words to encode hidden digits while using n-gram frequency distortion as the measure of detectability of the insertions. Using the publicly accessible Stanford Sentiment Analysis dataset of arpproximately 240,300 Tweets, we implemented the tweet steganization scheme -- a codeword (an existing word in the data set) inserted in random positions in random existing tweets to find the tweet that has the least possible n-gram distortion. We argue that this approximates Kullback-Liebler distance in a localized manner at low cost and thus we get a linguistic steganography scheme that is both formal and practical and permits a tradeoff between codeword density and detectability of the covert message. The extended abstract (8 pages) is available at the following link: https://arxiv.org/abs/2312.16840

[1] This is an extended description of a poster of the same title that was presented at the 2020 Information Theory and Applications Workshop held in San Diego, CA.
[2] Presenting author