### Background in Watermarking

# Explanation of watermarking in other types of media

Digital watermarking is a technology that enables embedding of information into multimedia (e.g. image, video, audio) without degrading the original fidelity of the content for copyright protection and content authentication. Watermarking in image and video contents has been extensively explored pre-deep learning (ref) and with the advent of deep neural networks, a new field of deep watermarking has also emerged (ref.).

# Transition to watermarking in NLP.

On the other hand, research in natural language watermarking is still at its infancy with only a handful of works in both pre- and post- deep learning era. Some challenges in embedding information into text while maintaining its fidelity are rooted from the discrete nature of text and the relatively low perceptual capacity of individual unit of text, which are words as opposed to pixels in image.

# Application of nlp watermarking and its necessity

Meanwhile, the recent proliferation of natural language content (such as …) has spurred the illegal piracy of such contents, which financially damages the original creators of the content.

(For instance, it is estimated that the web novel market has grown ).

In addition, the leakage of confidential or commercially valuable materials such as government documents and financial reports can also cause serious harm.

This calls for a secure way to guarantee copyright protection through ownership identification or leakage tracing through embedding secret messages into the text content.

# Emphasis on two key aspects of watermarking: payload and robustness.

### Prior works

# Previous works focusing in texts In image forms and their limitations.

Early works in text watermarking has focused on texts in image forms through controlling spaces between words and lines, which lends themselves defenseless to manual copy-writing of contents.

# Recent use of deep learning techniques. Detailed explanation of the two works (training based, algorithmic approach).

Recently, AWT proposed using transformer model to embed and extract messages into texts. This form of encoder-decoder for watermarking is a natural extension of methods in the image domain. While this showed the first feasibility of using neural networks for natural language watermarking, the quality of the watermarked texts was deteriorated due to relying entirely on the neural network without much constraint.

ContextLS circumvents this issue by relying on the neural network for lexical substitution while using rule-based algorithmic approach for the embedding and extraction of the watermarks.

# limitations of prior works

However, ContextLS does not take corruption of watermarked texts into consideration – either intentional from the adversary to hamper with the extraction process or unintentional in the process of leakage.

### Our contribution and approach

To build a robust watermarking system for natural language, we take inspiration from a well-known proposition from a classic image watermarking: that watermarks should be embedded in the fundamental structural component of the multimedia. Then the adversary cannot easily hamper with the watermark through corruption, because the fundamental structure of the content should be modified, which makes the whole purpose of pirating futile, for the watermark to be affected.

In this work, we sought to find the fundamental component of natural language.

A fundamental component should have the following properties:

1. Semantically similar sentences should tend to have the same fundamental component.

2. The fundamental component will be invariant to corruption by construction.

# We propose and study various domains on semantic and grammatical levels.

# Propose a general framework to approaching nlp watermarking considering robustness

# Analyze sources of errors and show how simple solutions can indeed mitigate errors.

We do not claim that the domain studied in this work is the optimal approach. We pave way for future works to explore other effective domains and solutions following our general framework.

### Framework

Given text X, we can derive states {S}. From state S, we can embed the message and attain the watermarked sets sets {X^{tilde}},

If we can attain the same states from the watermarked, then we can extract the message.

Attaining the states from natural language and the embedding state (though not exclusively) determine the robustness and payload of the approach.

For AWT, attaining the states and embedding is all done in a blackbox manner.

For context-ls, states are sequentially attained by those that pass the algorithm to ensure the message can be extracted without error.

For ours, states are determined as a function of the fundamental components.