Lumen: a Machine Learning Framework to Expose Influence Cues in Texts

UF UNIVERSITY of FLORIDA

Challenge:

- Expose influence cues (e.g., authority, scarcity) used in deceptive online texts such as phishing and disinformation to increase their appeal to the recipients.

Solution:

- Lumen, a machine learning based framework that exposes influence cues in text using a novel combination of well-known existing methods.
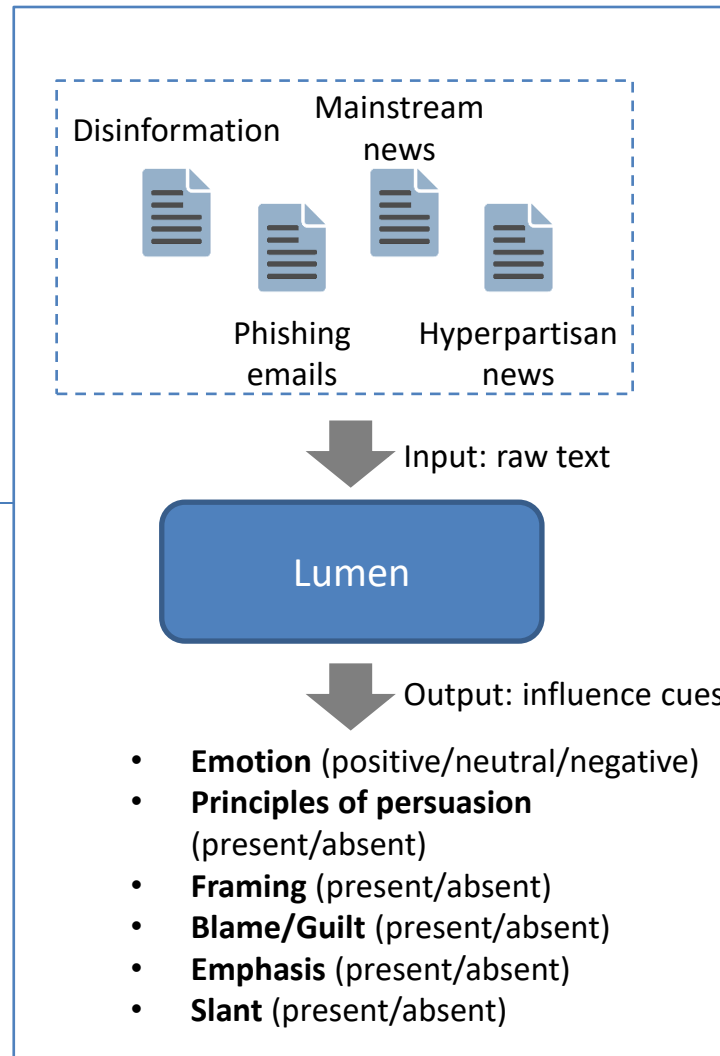- A newly developed dataset of 3K pieces of deceptive text used to train Lumen.

Disinformation

Mainstream news

Phishing emails

Hyperpartisan news

Input: raw text

Lumen

Output: influence cues

- **Emotion** (positive/neutral/negative)
- **Principles of persuasion** (present/absent)
- **Framing** (present/absent)
- **Blame/Guilt** (present/absent)
- **Emphasis** (present/absent)
- **Slant** (present/absent)

Scientific Impact:

- Competitive results than existing approaches but better interpretability.
- Findings may contribute to improve the accuracy of human-based detection of cyber-social engineering threats.
- One journal paper (under review).
- The dataset used to train Lumen is publicly available.

Broader Impact and Broader Participation:

- Can potentially benefit our society by triggering users to think slowly and analytically in the presence of influence cues, reducing their likelihood of falling for deceptive online content.
- Can potentially translate into practice through automatic labeling tools to assist human-based detection.