Toward Robust Annotation and Detection of Logical Fallacies in Online Misinformation

- Kenneth Huang, College of Information Sciences and Technology
- Dongwon Lee, College of Information Sciences and Technology

Abstract: Logical fallacies are reasoning errors that undermine the validity of arguments. They degrade online discussions, making arguments seem more dubious and contributing to the spread of misinformation. This proposal aims to combat misinformation and foster a healthier online discourse by detecting logical fallacies in online narratives, particularly in social media posts. A key challenge in logical fallacy detection is the low inter-annotator agreement (IAA), as even expert annotators often disagree on judgments. To address this, we will (i) develop an annotation scheme that accommodates multiple interpretations of logical fallacies within a single text instance, (ii) leverage this scheme to achieve a higher IAA for social media misinformation posts, and (iii) apply fallacy detection models to our proposed framework. Ultimately, this work aims to enable early interventions for misinformation by identifying when individuals begin to formulate fallacious arguments, which can lead to disseminating widespread misinformation.