Evan Li

Laidlaw Research Proposal

In today's social media saturated world, many are relying less and less on rigorously reviewed news sources for information. Instead, their world-views are shaped by content posted by other social media users. When users post about controversial topics, debates often ensue in the comment section. Those who passively read these debates are inevitably influenced by them, potentially leading to the spread of inaccurate information. Thus, before someone starts consuming or engaging in a particular discussion, they should be aware of possibly unhealthy aspects of that discussion.

I would like to research the presence of disinformation on social media discussion threads, where posts and comments are often informal and rarely verified for accuracy. I believe the presence of certain features of a discussion—lack of diverse viewpoints, extreme sentiment, ad homs, strawman arguments, and more—can act as warning signals for the presence of misinformation. For example, if a discussion thread is an echo chamber where all participants unconditionally agree with one another, misinformation can easily slip in without resistance. On the flip-side, a productive discussion will feature respectful disagreement, providing checks against false information.

I hope to apply my background in computer science and artificial intelligence to develop algorithms that automatically identify the quality, richness, and efficacy of a particular discussion thread on social media.

The health/productivity of a conversation is a vague concept. So, before the actual programming begins, I will research different metrics that measure the productivity of discourse, ensuring these metrics are as objective as possible and free from bias.

The latest advancements in natural language processing, a subfield of artificial intelligence, allow for the automatic detection of features in arguments such as similarity, stance, topic, and more. With these tools, various aspects of online conversations can be measured. For example, an NLP algorithm can be trained to detect if a conversation contains extreme antagonistic sentiment. These algorithms can be combined to provide a holistic account of the health of a conversation.

I believe this algorithm may provide valuable insights on the health of many salient online conversations that play a heavy role in influencing the viewpoints of millions. By collecting thousands of conversation samples from many online communities, I hope to compare the productivity of discourse in these communities, providing a general picture of the state of discourse on social media.

I also hope to see if there is a correlation between conversational health and disinformation. If a correlation is detected, conversational health may act as an effective warning signal for potential disinformation.