Samiha Samrose
samiha.mumu@gmail.com

Research Proposal

Problem Statement:

With the proliferation of social media (e.g., Twitter, Facebook, Reddit etc.), people are establishing network over online medium. Even though these communications are becoming a vital part of modern life, the regulation are still not equally strong. An evident of this is the #GamerGate incident [1]. Recently in the #MeToo incident [2], many of the tweets were mocking the movement. Cyberbullying and threatening have caused depression and suicide [3, 4]. Thus, hate speech in social media is becoming a serious issue. Another issue is - when we identify such a hate speech, what the approach should be to tackle it. With these premises in mind, I propose investigating the following research questions:

RQ1: Can automation identify or predict hate speech (example categories: bullying, threat, offense etc.)?
RQ2: Can virtual assistants be used to provide feedback to abusers in order to reduce online hate speech?

To exemplify, let us assume a hate speech expresses death threat to a particular person. In that case, the identification can be time-sensitive and the identification is thus crucial. Also, as this is sensitive, what should the language of the virtual agent be? If someone tries to post something offensive, besides identification how can the agent provide a feedback about the comment being insensitive? To answer these questions, the steps go as follows -

Research Plan and Steps:

1. Formulate examples of different categories of hate speech and run a study to label the data
   Or,
   Use an already existing labeled dataset
2. Build models to predict high and low sensitive hate speech
    a. Understand the characteristic of the data
        i. Sentiment analysis (vader/nltk toolkit)
        ii. Language analysis (liwc toolkit)
    b. Deploy machine learning tools
        i. Classification/Clustering (scikit-learn, pytorch)
3. Design a virtual assistant to provide feedback
    a. Design a survey to perform requirement analysis & prototype building
    b. Get rating from user on different setting (e.g., language of agent)
    c. Get rating from user on different design (e.g., chat vs popup)

For data space, I want to look into text data from social media (e.g., Twitter). In my current research, I have worked with analyzing the characteristics of #MeToo contents. Alongside, my primary research work is improving group interactions in a video-conferencing based platform integrated with real-time and post-discussion feedback through chat-based virtual assistant. In the proposed research questions RQ1 and RQ2, I would apply my previous knowledge to reach the objectives.

Requirements, Risks and Backup Plans:

The resource requirements would be the online libraries for coding, data labelling, thorough user study. However, there are risks involved too, such as –

1. The data categories have to be well thought and precise. Otherwise the labels would be less useful.
   Backup plan: Follow standard category pattern.
2. Data labelling may require long amount of time.
   Backup plan: Use already available labelled data.
3. Classification may not achieve significant accuracy.
   Backup plan: Test out multiple small sample cases.
4. Virtual assistant design may not achieve enough System Usability Score (SUS).
   Backup plan: Have low fidelity prototype for primary testing
5. Timeline may not be sufficient to answer both RQ1 and RQ2
   Backup plan: Prioritize a research question, or divide work.

Future Impact:

The solution can be useful in further analyzing hate speech from different perspective. For example, this model can be extended for audio based conversations. The model can be used on the audio transcript. The virtual assistant can be implemented not only as a chatbot but also as a voice agent too. These future applications show how the model can be generalized for various cases. However, for actual impact, these research question can address the very serious issue in online community. This work can assist in ensuring a safe and respectful environment in online social network.

References:

[1] https://en.wikipedia.org/wiki/Gamergate_controversy
[2] https://metoomvmt.org/
[3] https://www.meganmeierfoundation.org/suicide-statistics.html
[4] https://www.meganmeierfoundation.org/cyberbullying-social-media.html