SLIDE 1: Hello CSER people! I'm Montgomery Gole and am here to present my research on Sentiment within the newcomer Linux Kernel Development Community. I am currently a 4th year UofT Faculty of Information student studying for my Bachelor of Information, and was given the opportunity to work with Dr. Kelly Lyons to get a taste for the process of conducting academic research.

SLIDE 2: Taking inspiration from both Zagalsky, et al.'s How the R community creates and curates knowledge: an extended study of stack overflow and mailing lists, and Ferreira, et al.'s A longitudinal study on the maintainers' sentiment of a large scale open source ecosystem, I attempted to analyze sentiment from both the Newbie Linux Kernel Mailing List, and StackOverflow's linux-kernel tag.

SLIDE 3: Some things which motivated me to Study Sentiment within the Newbie Linux Kernel Community comes from the fact that sentiment can affect the efficiency and synergy within any software development team, especially one like the Linux Kernel Dev team which relies on volunteers. In fact, the paper "*The "Shut the f**k up" Phenomenon: Characterizing Incivility in Open Source Code Review Discussions*" found that ⅔ of non-technical emails within the Linux Kernel Mailing List included uncivil, negative attributes. While this paper was released after I concluded my research it captures the essence of what I attempted to find.

SLIDE 4: These motivations eventually led me to my research question: what is the sentiment like within newcomer spaces of the Linux Kernel Development community?

Slide 5: I collected data from January of 2019-December 31st 2020, this was partly to see if the stressful time of the onset of the COVID-19 pandemic would have any effect on the sentiment.I collected data from the Newbie Linux Kernel Mailing List, utilising bash script with wget to download data, and gzip to extract every email from the archive within my date range. Also, instead of downloading data from the internet archive like in Zagalsky et al.'s replication package, I decided to collect data from the Stack Exchange Data explorer, where I could use SQL queries to select specific data from any StackExchange Site. I also parsed the data with Python and a lot of Regular expressions.

Slide 6: Some things I found from SO WERE that the posts per month actually goes up after covid hit, and that a majority of users have a reputation between 10 and 100 on the linux-kernel tag. In the noob mailing list every person who asked a question was counted as a noob. I also created a more indepth typology of the SO users, numerically defining newbies with 0 accepted answers, an intermediate developer was defined to have an accepted answer count z score between 0 and 2, an expert would be in at least the 98th percentile of users with respect to their accepted answer count, and a z-score greater than 2. With the Linux Kernel Mailing List, I had a tougher time, as the parsing was complex; I created a python script to do so. In hindsight a python module like pandas would have made this task more computationally efficient, and easier to be viewed analytically.

Slide 8: Taking from Ferrera et al.'s method, I used Senti4SD to analyze the data I had collected, and parsed. I found that newbie Stackoverflow posts are mainly neutral, but there are more negative posts than positive posts.

Slide 9: The next steps of my research will include performing a sentiment analysis on the LKML. I would also continue with writing my paper, and eventually at least have a final product which could hopefully be published.

Thanks so much for listening, everyone. Are there any questions you'd like to ask about my research?

If anyones recruiting grad students I'm in the process of applying to grad school for Computer Science!!!