Comparative Analysis between TextRank and LexRank: Testing Perceived Utility of Text Summarization Models for Data Annotators

Avery Reyna
University of Central Florida
avery.reyna16@knights.ucf.edu

Ashwaq Soubai
University of Central Florida
atalsoubai@knights.ucf.edu

Pamela Wisniewski
University of Central Florida
pamela.wisniewski@ucf.edu

Motivation and Big Idea

- Text summarization models have been widely used to condense large pieces of text to shorter ones while preserving key informational elements and the meaning of content.
- While this has been used in a variety of settings, such as crafting news article summaries and encoding deep learning algorithms, this project aims to utilize these text summarization models in the research setting itself for those that must annotate large sets of textual data.
- Specifically, we want to analyze the outputs of two different types of summarization models to discover which one has better utility.

Research Questions

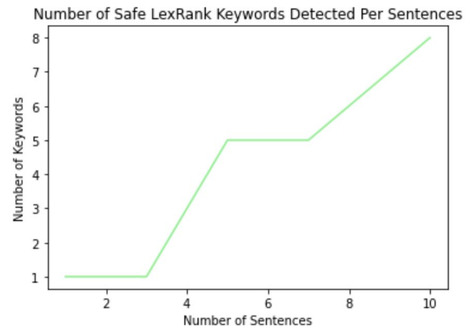
- Which type of text summarization model should Data Annotators utilize when dealing with extensive textual data?

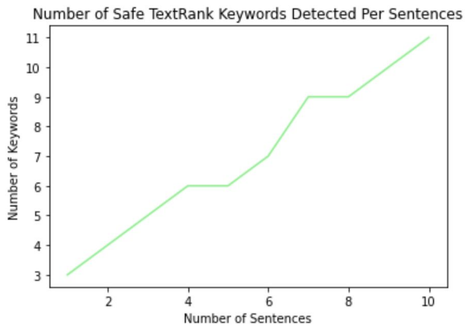
Methodology

- Utilizing the programming language of Python and its associated Natural Language Processing (NLP) packages to access the automatic text summarization models of TextRank and LexRank.
- We applied these models on (N = 20) conversations, with 10 of these conversations containing unsafe language and the other 10 having safe language.

Findings

- Due to the nature of the conversation data not always following proper sentence structure and maintaining a high level of coherence, we started to expect that the outputs of the summarization models would either contain some sentences or an agglomeration of words from the larger set of text.
- Come to find out, we were getting sentences from the conversations being summarized and upon further inspection, the sentences that were outputted contained keywords that signal whether a certain conversation is safe or unsafe.
- This is specially useful for data annotation because if researchers can quickly input a large amount of text into a summarization model and receive an output that contains sentences with important keywords, the time spent sifting through each sentence is substantially decreased.





Discussion

- After discovering that these text summarization models were outputting sentences that contained keywords that signal if the inputted conversations were safe or unsafe, there are three more steps to complete our analysis:

Better understand and define what keywords are and assess the outputs from TextRank and LexRank from that lens

↓

Complete the interview process with all the Data Annotators in order to have an extra layer of qualitative analysis

↓

Figure out if there are any differences in the outputs when unsafe conversations are inputted

Acknowledgements

This research is partially supported by the U.S. National Science Foundation under grants IIP-1827700. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the research sponsor.