Proposal

Magdaleno, Alejandro
amagale@student.kennesaw.edu

Lyles-Woods, Quin'darius
qlyleswo@student.kennesaw.edu

October 17, 2021

Proposed Study

For the topic of study we have decided to venture into the arena of political language. More specifically we are looking to understand more about the voice of prose and seeing if that can be algorithmically detected with Natural Language Processing. We will build a classification model to determine what news outlet a given body of text has originated from.

Goals

The goals of the proposed study to build a classification model to determine where bodies of text originate from will be first and for most to be able to, of course, detect where the bodies of text originated from. An auxiliary goal will be to tell which of the author has written the given body of text. Another auxiliary goal will be the level of accuracy the model will have. Aiming to be better the 70% to begin with and gradually increasing the threshold. If we pass the 95% mark we will considered the first auxiliary goal mentioned which is to classify the authors voice in a given piece.

Relevance

The relevance of the proposed study of algorithmically detecting the origin of new articles can be linked to the analysis of bias and unearthing it within the large repertoire that we are given with our selection of where to find and digest news of the world around us. If we can find out the bias this should be the first step and shifting through to the truth in these sometimes lengthy articles that share more about the authors voice than the subject matter at hand. Doing this will save a great deal for anyone that just wants the facts fast and will allow those individuals to regain parts of their lives that would have been ultimately lost other wise.

Data

For this project, we want variety in the data that we extract in order to detect the best patterns in the data. Our data will consists of many different transcripts from several targeted news stations on top of news articles, and posssible transcribed voice recordings of news broadcasts. We could be set with just using the transcriptions already provided on news station websites but variety is key in data and bringing in articles and captioned voice recordings should ideally start giving the model more to go off of in terms of patterns that it can detect. There won't be much need for data modification in this case besides taking in voice recording and captioning them and many different news station websites already provide their own transcriptions. One example is the website,

https://transcripts.cnn.com/, where cnn provides all their show transcripts online and many other stations also provide this.

Models and Algorithms

The data will need to be prepared first. The data consisting of the transcriptions will also include a label for it's output. A possible use for an encoder can be used for the target output value in order to categorize the outputs. After the data is organized and set, we will also need to use get some features from the data such as getting context through TF-IDF, topic features through context, or other methods like count vectors. After some initial features are received from the data, this is where model training will start to kick in. There are some routes that can be taken here from using traditional artificial intelligence strategies like naive bayes classifier, or going into deep learning models such as convolutional neural networks(CNN). The use of deep learning could serve a great advantage here because we can have the use of non-linearity. A deep learning model such as a CNN can take find and extract more patterns through training and eventually find the best model to solve the problem. In the case that a CNN is used. The first layers in the network would be used to embed the sentences into low dimensional vectors. Then we would run multiple convolutions or filters on the input data in order to create a feature map of the best patterns that are being found on the inputs. This will eventually start leading us to more accurate predictions. Afterwards, a max-pooling of the layers will come in which will attempt to calculate the maximum value in the feature maps from the previously filters layers. At the end of the model, a softmax classifier can be reached to give a target news station that the data came from. This is one example of an ideal model that can be used for our project.