ANLY580 Project Proposal
Due: Nov 10, 2021
Final Project Proposal
Gabriella Zakrocki and Scott Johnson

For our ANLY580 Final Project we intend on creating an algorithm that classifies news articles into a range of category descriptions. Specifically, this algorithm will be constructed as a neural network that will classify a news article based on the vocabulary present in the headline. To train and test this neural network, we intend on utilizing data from an open source Kaggle dataset. This dataset includes approximately 200K news article headlines already labeled with their respective category description. Within the data itself there are a total of 41 different class labels (a list of these categories and corresponding article counts can be seen in Table 1 of the Appendix). Additionally, this data also includes the author, date, and link to each article.

Our motivation for this project lies in the specificity of classification that could be achieved through using a high number of labeled categories. During our initial research, we observed multiple datasets like ours, that contained labeled news article headlines, however most of these datasets only included up to four different category labels. Moreover, most of these dataset's labels were quite broad and already distinct from each other, for example: Business, Sports, Entertainment, and Politics. Thus, any model we created from such datasets would be limited in its ability to provide additional and beneficial insights to an individual; as it is reasonable to assume that a single human could ascertain which of those four named categories (Business, Sports, Entertainment, and Politics) a news article fell into relatively quickly on their own. Furthermore, we recognize that there are so many additional topics that are covered in the news, and thus these four labels did not allow for full inclusion of the range of topics discussed and written about in life. Because of these factors, we are motivated to construct a model that classifies news articles into a larger range of categories, as this is more reflective of the real world.

Although we hope to utilize a larger number of labels to reflect a greater proportion of the topics news articles cover, we recognize that this in turn raises the complexity of our model. With this in mind, we have decided that although our data includes up to 41 different label categories, we will not be utilizing every label in the creation of our model. This decision has been made to alleviate concerns regarding the extreme model complexity, and subsequent computational power and time necessary, to succeed in creating our model. Even so, although we will not be including all 41 different label categories, the exact number of labels we intend on utilizing is still unknown. Currently, we believe that we will not use more than 10 different label categories, however we recognize that even 10 different label categories may prove to still be too complex of a model. Because of this, we intend on first trying to classify the news articles into 8 different label categories and evaluating the success of that model. Depending on the results of that model, we will either increase or decrease our number of labels as necessary to create a

sufficient model. Overall, we perceive that one of our biggest challenges in this project will be balancing model complexity and successful classification into a greater range of categories as desired.

To evaluate our prediction model we will determine the accuracy rate on the test set. Furthermore, varying the amount of layers, batch sizes, and epochs can additionally help to determine our optimal model that results in the best accuracy rate. This accuracy rate will then inform how well our model will correctly predict the category of unlabeled news articles.

In conclusion, our goal is to create a neural network that will classify news articles based on their headline into between 8 to 10 label categories to better reflect the vast variety of news shared in the world. We then intend on presenting the results of our undertaking in the form of a written report that discusses how the neural network was created, what labels were chosen from the data and why, as well as the evaluation metrics for our model. Furthermore, we will also provide a brief live demonstration of our model at work. We will do this by presenting the class with an unlabeled news article, have students read the headline, and ask them to determine in which label category would they classify the article. Following that, we will then input the unlabeled article into our model and observe the results of where it was in fact classified. We believe that through a written report and live demonstration we would be able to provide sufficient information regarding the model and our reasonings, while also providing a hands-on experience as to how the model actually works.

Appendix

Table 1: Data Label Categories and Corresponding Article Counts

POLITICS: 32739	HOME & LIVING: 4195	SCIENCE: 2178
WELLNESS: 17827	PARENTS: 3955	WORLD NEWS: 2177
ENTERTAINMENT: 16058	THE WORLDPOST: 3664	TASTE: 2096
	WEDDINGS: 3651	TECH: 2082
TRAVEL: 9887	WOMEN: 3490	MONEY: 1707
STYLE & BEAUTY: 9649	IMPACT: 3459	ARTS: 1509
PARENTING: 8677	DIVORCE: 3426	FIFTY: 1401
HEALTHY LIVING: 6694	CRIME: 3405	GOOD NEWS: 1398
QUEER VOICES: 6314	MEDIA: 2815	ARTS & CULTURE: 1339
FOOD & DRINK: 6226	WEIRD NEWS: 2670	ENVIRONMENT: 1323
BUSINESS: 5937	GREEN: 2622	COLLEGE: 1144
COMEDY: 5175	WORLDPOST: 2579	LATINO VOICES: 1129
SPORTS: 4884	RELIGION: 2556	CULTURE & ARTS: 1030
BLACK VOICES: 4528	STYLE: 2254	EDUCATION: 1004

Link to Data: https://www.kaggle.com/rmisra/news-category-dataset