HON 451 Thesis Proposal

My name is Hunter Berry, and I am a double major here at Berry, studying Political Science and Cyber-Law (an IDM). For my HON 451 thesis, I am studying the use of aspect-based sentiment analysis and machine learning for the purposes of determining political ideology amongst the American population. Essentially, my project is taking data from Twitter, specifically the tweets of U.S. politicians and celebrities with clear ideologies, and using their tweets to identify patterns that can in turn be used to determine the likelihood of any given person being from a certain political party or ideology when given that person’s Twitter account, regardless of whether they clearly express their ideology or not.

This semester, my research has focused on the initial aspects of creating this algorithm. For instance, one of the first goals of the project was to create an actual means of gathering and sorting the data. This included the process of getting an authorized Twitter academic development account and API key, gathering a list of all United States Congressional members and their respective Twitter accounts, and a data collection script. This script, which can be viewed in the project’s GitHub, gathers the latest tweets from each Congressional member’s twitter page, and loads the data into a Comma Separated Values (CSV) file. This file contains information including the party of the user in question, the state in which they are from, the overall “sentiment” of the tweet, the various noun phrases and terms used within the tweet, and much more. This file also uses the information gathered in these tweets to make tables and frequency charts detailing the number of uses of various terms and phrases by each party, in each state, and by each party in each state.

Furthermore, this semester’s work has also cumulated in the creation of a basic classification algorithm that can classify tweets by party and by state. Using some of the frequency charts that I’ve created, this classification system takes in the polarity of a given tweet and the terms and phrases within it to attempt to determine where the tweet came from or what party it may have been. Although this classification system is rudimentary in its current form, it will be improved upon over time.

Next semester, the goal of the project is two-fold, focusing on finishing the algorithm that was started this semester, and the creation of a paper. As for finishing the algorithm, although there is a basic classification system already in place, next semester will focus on how this classification system can be improved, through things like more honed algorithmic methods (such as Naïve Bayes, the XGBRegressor machine learning models, etc.), the use of more detailed information, such as the aforementioned term and phrase frequency tables, and more. As for the paper, this paper will not only highlight the algorithm, its creation, and its effectiveness, but the paper will also discuss the hypothetical implementation of the algorithm on both a domestic and an international level, as well as looking at similar algorithms and technologies, such as the [2017 Stanford “gaydar” AI](https://osf.io/zn79k/), and the impact said algorithms could have on the world.

Overall, my plan for HON 451 is to continue the hard work that I’ve put into my current algorithm and build on it so that it becomes much stronger. I’d like to continue studying the process and look into things like machine learning that will help create a more accurate algorithm, and show the potential affects of my own research and other topics like it.