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 political ideologies (i.e. conservatism, liberalism, centrism, etc.), 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 feed, 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.

Many of these aspects were made available to me from a number of Python libraries, the most used of which was the TextBlob library. This library is more user-friendly and simplified version of the Natural Language Tool Kit, which provides language analysis tools for those programming in Python. The TextBlob library has many functions, although two I have already began making use of is the library’s sentiment/polarity analyzer and noun phrase tagger/extractor. To use the library, I can give the library any text, in this case an excert from someone’s Twitter feed, and have it return the noun phrases that were within that tweet, and the sentiment of the tweet, which is calculated by the library by looking at the sentiment of individual phrases and terms, and the sentence’s structure and goal as a whole. As previously mentioned, this allows me to begin analyzing the tweets and searching for patterns, i.e. determining if one party or one state is more likely to use a certain term or phrase or have a certain sentiment range. For instance, in early testing, I found that politicians with a liberal bias were much more likely to use words like “environment” than were those with a conservative bias (this was a roughly 225:1 ratio).

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 a variety of different factors, such as the frequency of terms and the party/state they may have been corelated to, and uses that to determine where a given tweet may have originated from or what party it may be associated with. Although this classification system is rudimentary in its current form, it will be improved upon over time. The current algorithm in place was able to predict around with around 70% accuracy when given ~2000 lines of data. My belief is that when further tuning is done, more models are tested, and more data is gathered, a classification algorithm may be able to predict parties and states with upwards of 90% accuracy.

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, support vector machines (SVMs), the XGBRegressor machine learning models, etc.), the use of more detailed information, such as the aforementioned term and phrase frequency tables, and more. As of now, the algorithm is based mostly on information derived from the part-of-speech tagger and noun phrase extraction system. However, my goal next semester is to continue on some of the research I’ve done to build a more hybrid algorithm that encompass sentiment analysis and polarity as well. As was mentioned by a number of sources, a hybrid-approach combining a lexiconic/sentiment-based analysis and a machine learning based analysis may provide the best results (see Dandrea, Alessia, et al, Liu, Bing, and Hasan, Ali).

Furthermore, I’d like to look more into aspect-based sentiment analysis. This process, as described in Liu’s book, breaks/categories tweets into different ‘genres,’ in the case of this project, things like economics, the environment, the court system, and more. It then allocates each of these genres a sentiment or polarity. Coupled with things like hashtags, this would be an excellent way to provide data to a machine-learning model and create a hybrid model that uses individual sentiments about different genres and the phrases/terms used to determine the party/ideology a tweet may best relate to. I hope to build on the part-of-speech tagger and noun phrase extractor from the TextBlob library to create a sub-classifier that would be able to take a tweet and give some of the categories it may belong to, perhaps using a Hidden Markov Chain or Conditional Random Field (CRF) model to help me do so.

As for the paper, this paper will begin by highlighting the algorithm, its creation, and its effectiveness. I hope to discuss the algorithm as I currently intend for it to work – a program that, given the username of a Twitter account, takes any number of tweets made from the said account, and determines the ideology that the said person most likely follows. It does this by breaking the user’s feed apart to determine each individual tweet’s overall sentiment and aspect-based sentiment (i.e. the user’s sentiment/polarity on the economy, on the environment, on the presidency, etc.), and combining it with the overall frequency of terms and phrases used within the feed. Coupled with information like the state the user came from (which may come from Twitter, but can also be partially pieced together by a separate Naïve Bayes classification algorithm), the algorithm will hopefully output a party based on a comparison to the data that was gathered from politicians and celebrities. After discussing this, my goal is for the paper to 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. Since technology plays such a prominent role in today’s society and can be used for both good and bad depending on the hands it is in, I hope to show how a more sophisticated version of the algorithm could change the world we live for better or worse.

Overall, my plan for HON 451 is to continue the 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 effects of my own research and other topics like it.