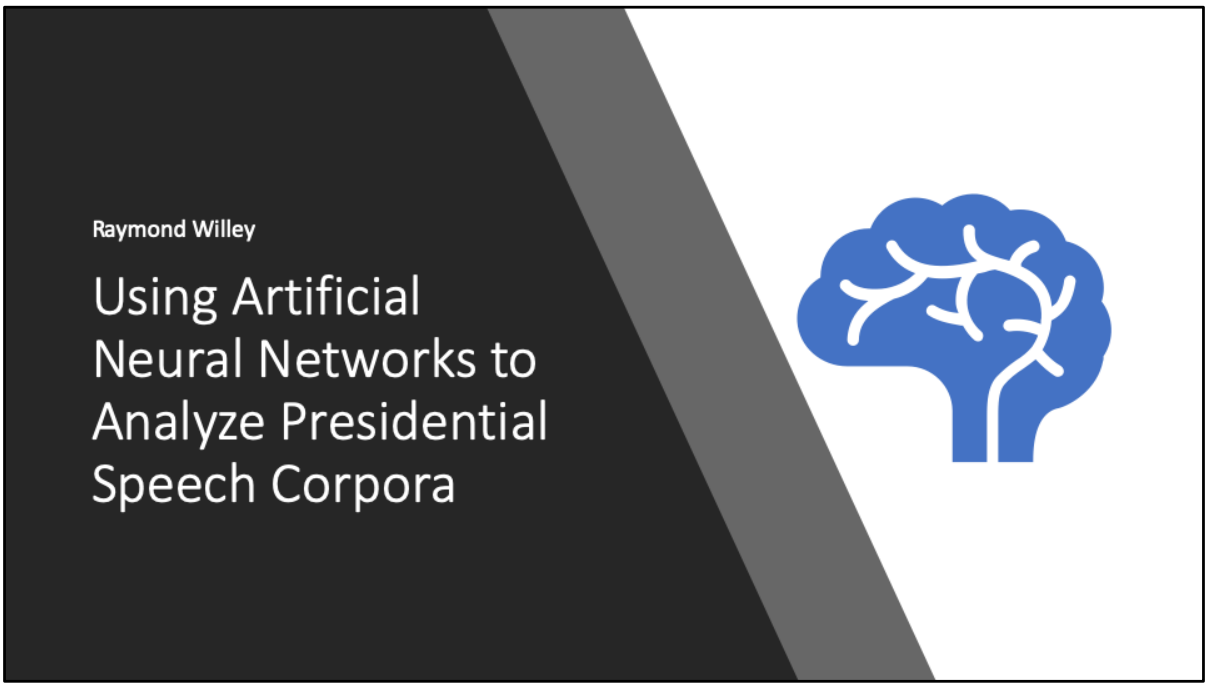
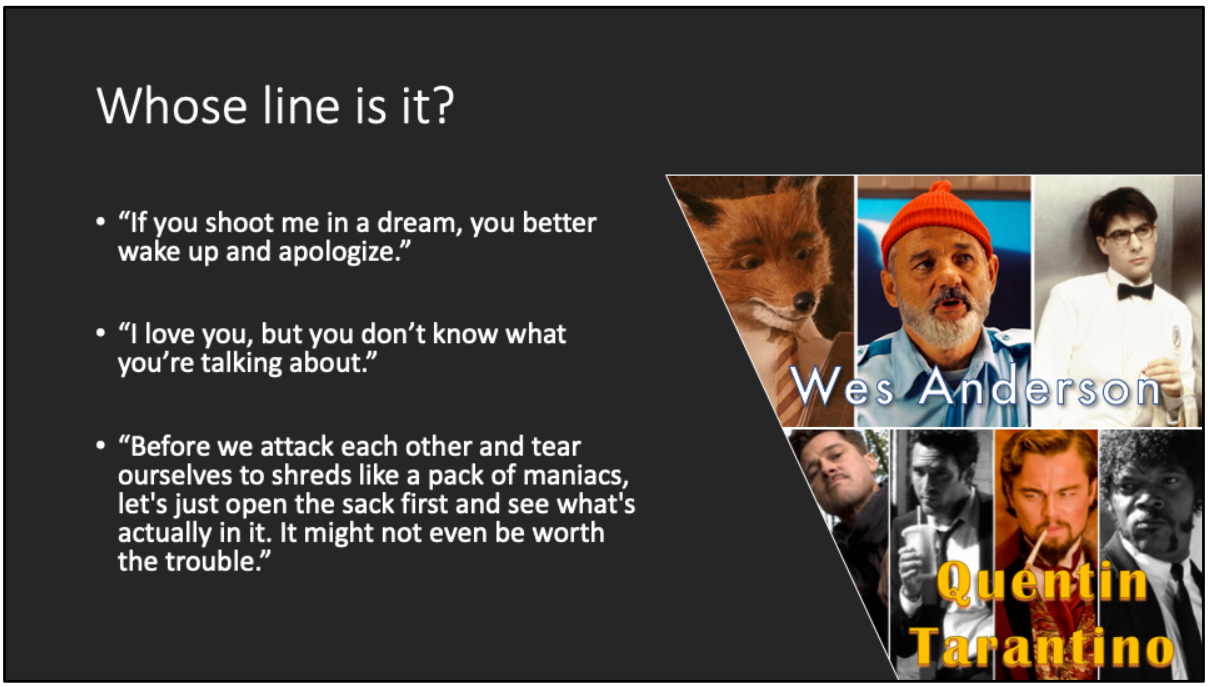


Note: For a video overview of this presentation, visit www.overloadblog.info. The video will include a brief overview of how neural networks operate, and how they are often applied to NLP tasks.

Hello everyone. Today we are going to be looking at how artificial neural networks can be used to analyze presidential speeches.

To understand the nature of the objective here, let's start by playing a game.



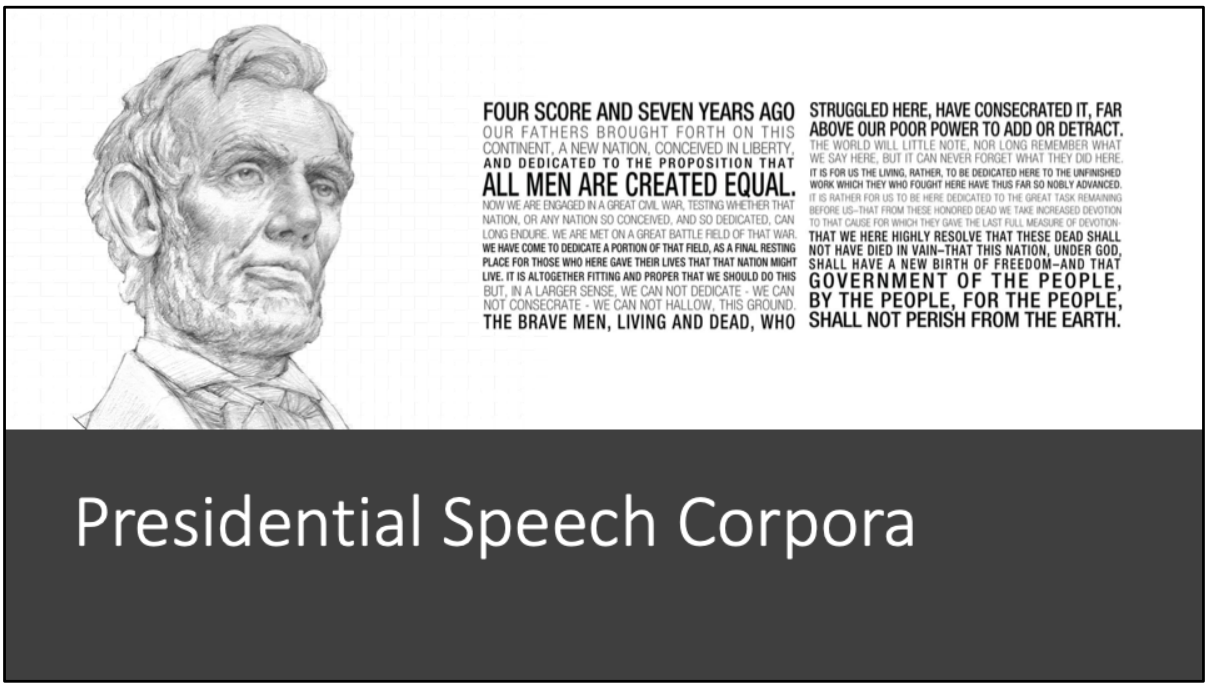
Consider two well-known, yet very distinct filmmakers: Quentin Tarantino and Wes Anderson. For simplicity sake, let's say that movies by the former might be described as *violent,* while the latter would be described as *quirky*. Armed with this knowledge, see if you can correctly identify the writer of each of these movie quotes.

Here, we see that the first two quotes are pretty easy to identify. The first is from the Tarantino classic *Reservoir Dogs,* while the second is from Anderson's *Moonrise Kingdom.* The third one, however, is a bit more challenging in that it's both a bit violent and bit quirky.

My wife, who is a big fan of both writer/directors, hadn't yet seen this movie, and guessed that it was from Tarantino. In fact, it is from Anderson's latest film, *Isle of Dogs.* What is interesting about this is that while the first two quotes make it easy to see the differences between them, the third highlights some of their similarities.

Given a series of quotes like this, chances are you'd get more than a few wrong. But those mistakes can actually help us better understand what two distinct individuals have in common, when it may not be obvious at first glance.

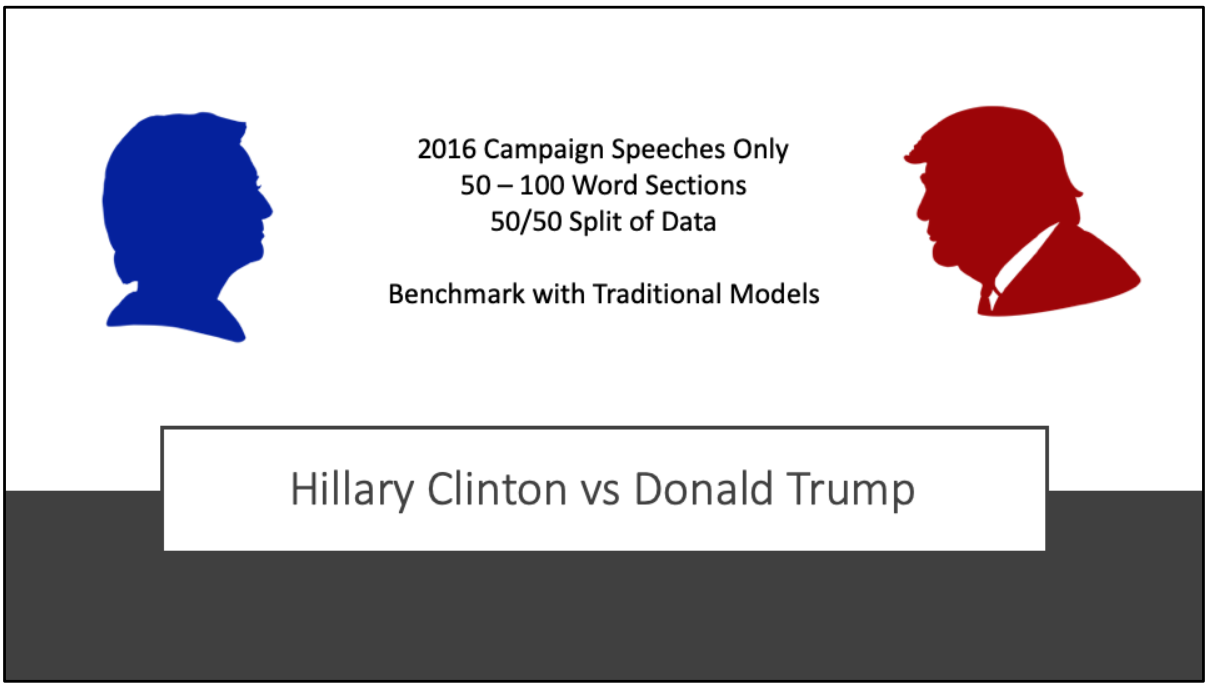
The challenge we have is to see if we can build and train an artificial network to do the same thing and provide the same type of insight. Only, instead of looking at movie quotes…



Presidential Speech Corpora

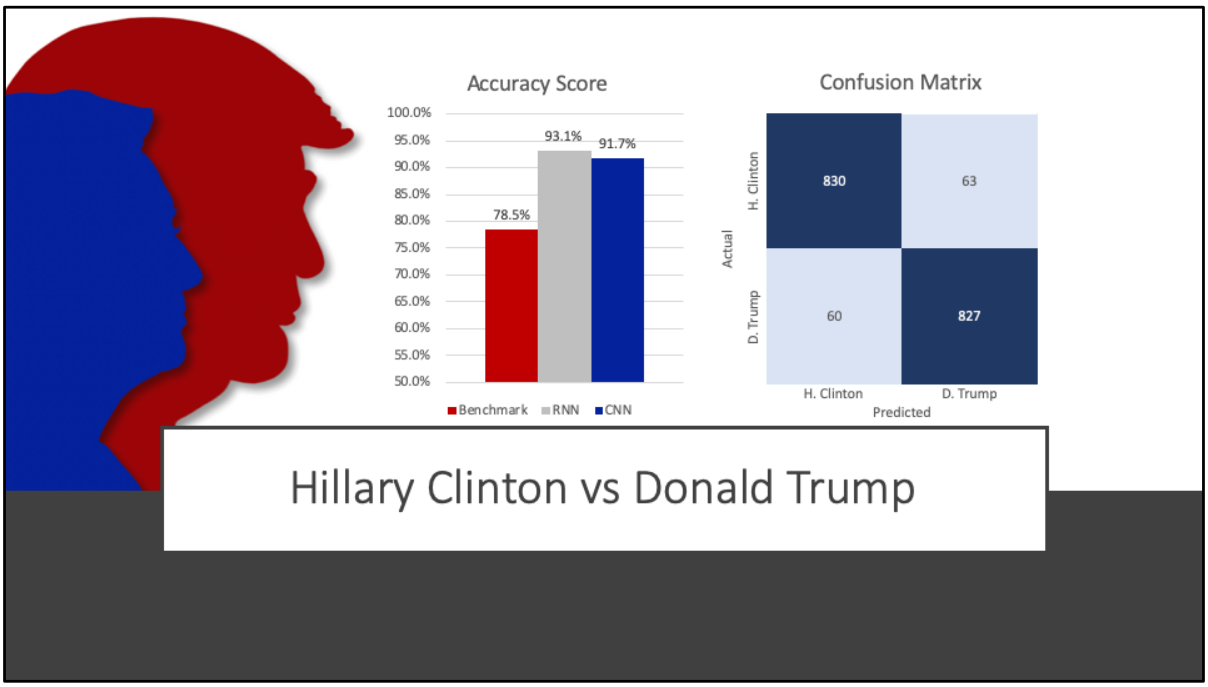
…our network is going to learn from presidential speeches. Can a network read a subset of passages from two presidents' speech corpora, and use them to correctly identify the speaker of passages it has never seen before? Well, let's find out.

To start, we want to look at two individuals with whom most people have strong familiarity, and who are known to be near opposites. This will make it easier for us to interpret the qualitative aspects of the results, allowing us to derive better insight to the strengths and weaknesses of our network before applying to presidents who are less familiar.



Where better to start than by looking at our two most recent presidential candidates: Donald Trump and Hillary Clinton? We are only looking at speeches each candidate made on the 2016 campaign trail. Or network will take data in 50- to 100-word chunks, training on half of the data, then using what it has learned to correctly identify the rest.

To truly gauge how well our network performs, we want to establish a benchmark using traditional machine learning methods: because artificial neural networks are much more resource intensive, we should expect that they will at least outperform traditional models.

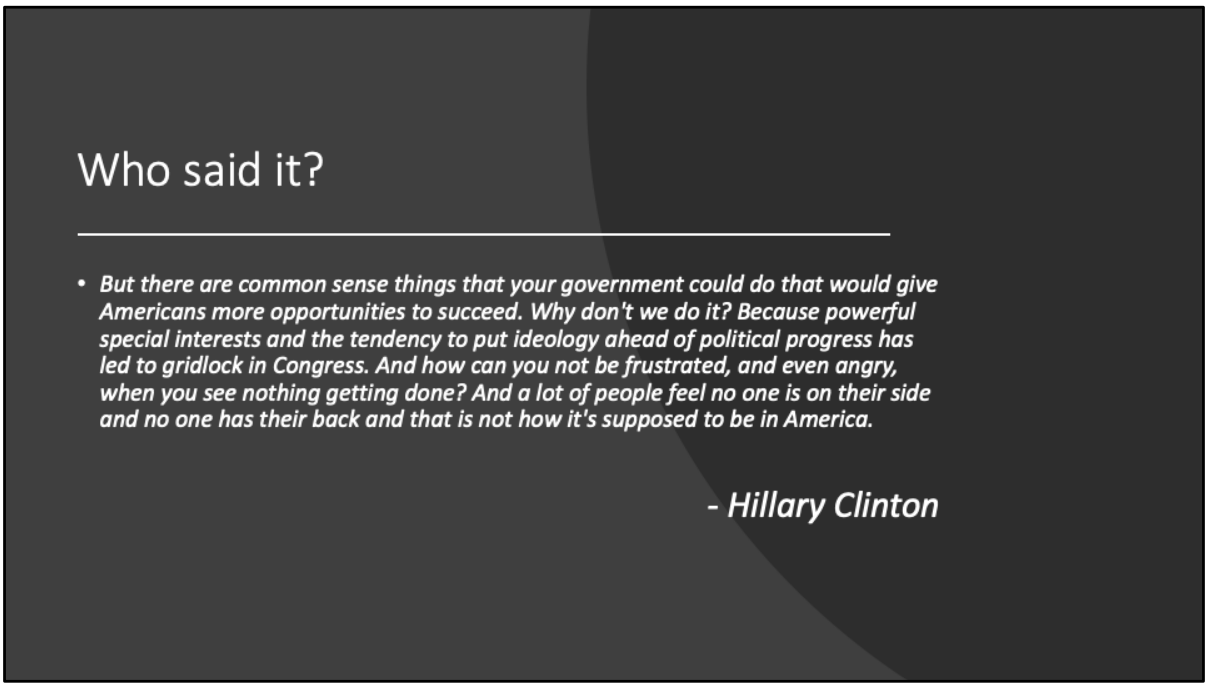


The best of three different model performances are chosen as a benchmark, and in this case, a logistic regression was able to achieve an accuracy score of 78.5%.

So, how do neural networks compare? Well, two general types of networks were trained and tested. The first is what's called a Recurrent Neural Network (RNN for short), which has an architecture that has been well optimized and is frequently used for these types of tasks. Sixteen different variations of this architecture were applied, and it was able to achieve an accuracy score of 93.1%! Not bad for a first run!

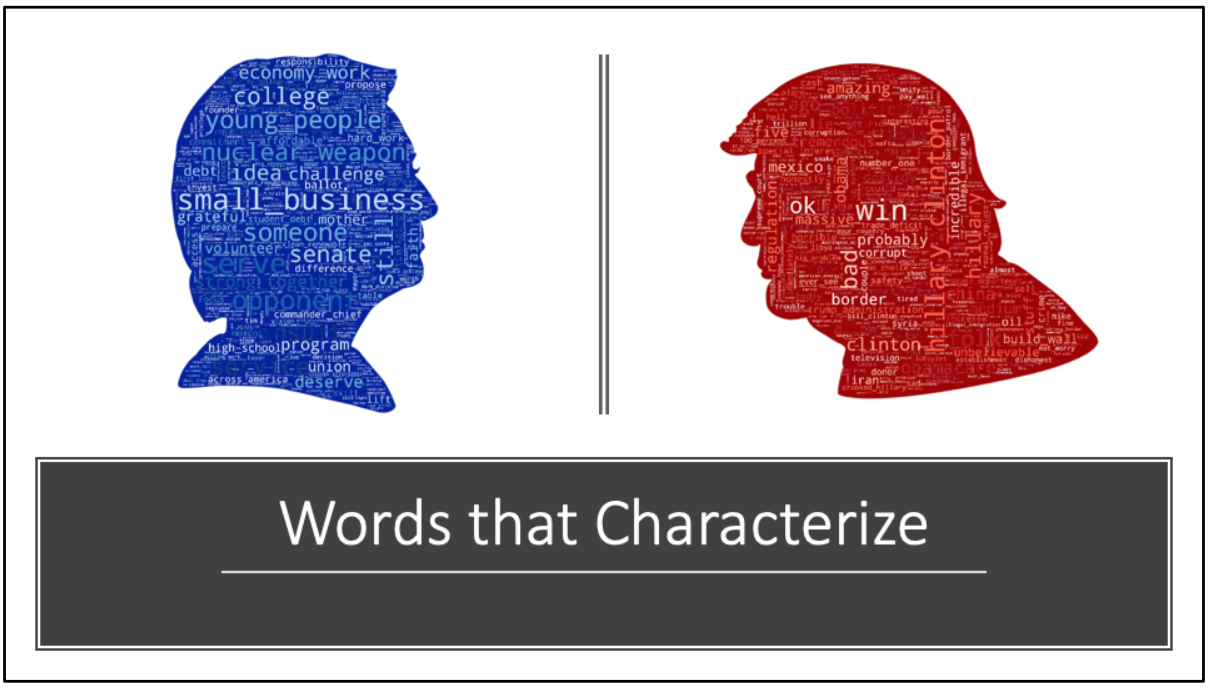
The other model applied was a Convolutional Neural Network (or CNN), which is traditionally used for image recognition, but has recently begun to show promise in the NLP space as well. These run considerably lighter than Recurrent networks, but the structure used also has more parameters that can be adjusted, so 72 variations of this model were trained and tested. So, how did our Convolutional network do? Well, not quite as well as the Recurrent network, but it still achieved a 91.7% accuracy score. So, both of our models demonstrated significant improvement over our benchmark, but for a deeper analysis of the results, we'll use the predictions made by our top performing model.

In total, 1,780 predictions were made, of which 123 were inaccurate. It's worth noting that there are two ways in which an error can be made: either by predicting that Trump said something that was actually Clinton, or that a passage was spoken by Clinton when it was in fact Trump. From the upper right and lower left quadrants of the chart, we can see that the division of these error types are well balanced. With that in mind, let's have a look at an example of a mistake that our network made, and see if you can correctly identify the speaker:



Given that Trump largely ran on a platform that was based on shaking things up in Washington, it wouldn't be far-fetched to guess that this quote came from him. But such a guess would be incorrect, as this one comes from Hillary Clinton.

This result may or may not surprise you, but the main takeaway here should be that this type of quote highlights an area where the two candidates seemingly have something in common. Still, an accuracy score of over 93% indicates that these candidates have a lot more that divides them than unifies them. So, let's see if we can get a sense of their differences by looking at their word usage.



Rather than looking at just the most common words for each candidate, we want to look at the words used more frequently by one candidate relative to the other. Starting with Hillary Clinton, we see frequent references to small business and young people, as well as references to service and senate, likely reflecting topics related to her experience in office.

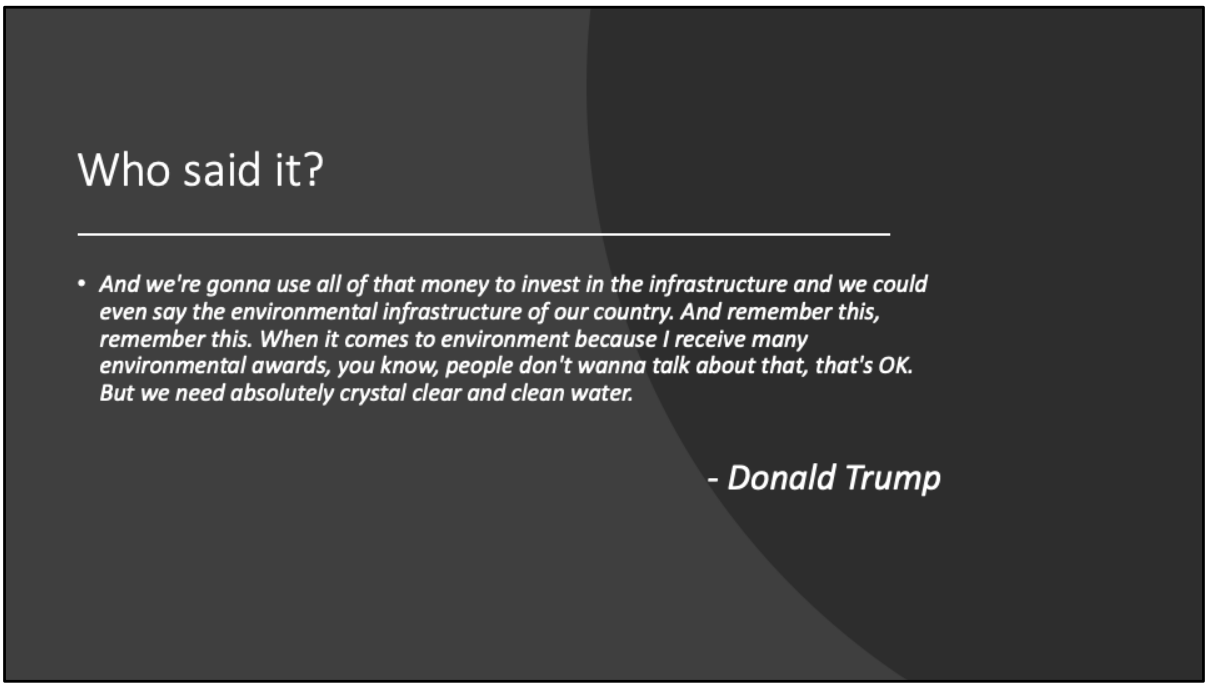
With regard to Trump, what jumps out the most here is that there are a lot of direct references to Clinton, more so than any other topic. While we do also see references to immigration, trade deficits, and regulations, their prominence is much less pronounced than one might expect.

However, there's one other set of words I think is worth looking at, and that is the most common words that exist within the space of prediction errors. Can we identify common themes that seem to be a bit more universal?

Themes within Errors

Here we can see certain words and phrases that are more or less par for the course for any politician running for president: American flag, work together, etc. Even the phrase nuclear weapon is going to be a prominent topic for any presidential campaign, so we want to look at the words that are less common.

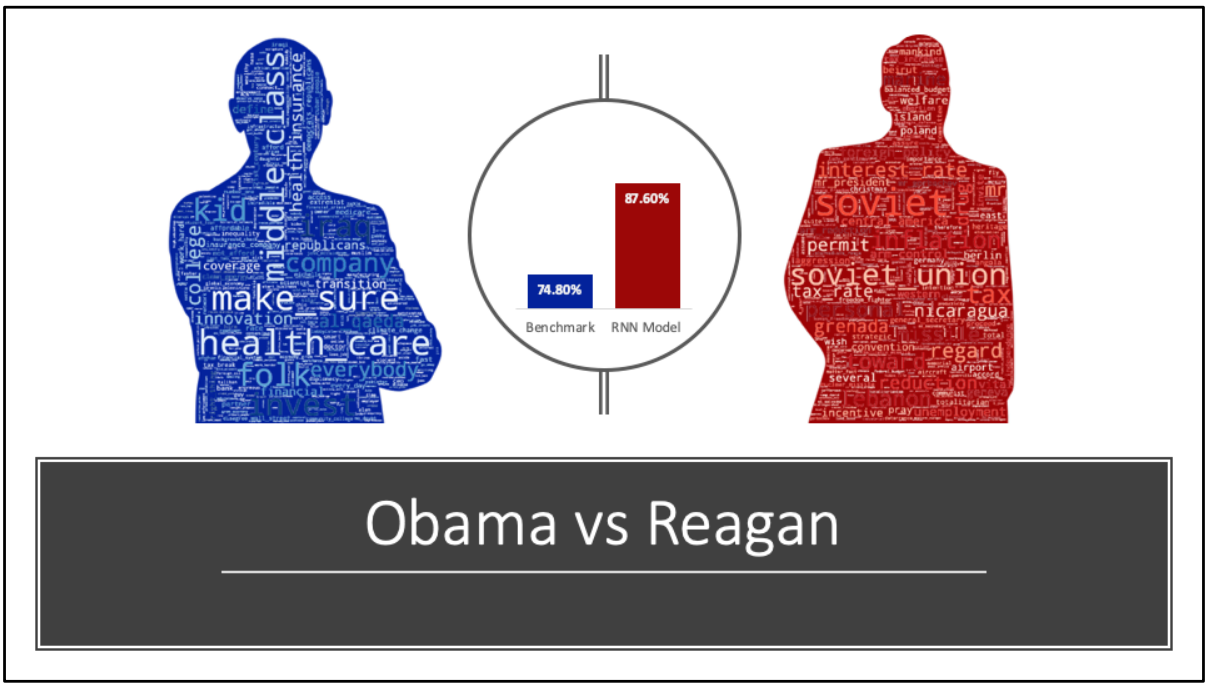
The word *opportunity* lead us to that quote from Hillary earlier, so let's try another one: *infrastructure*. Here's a quote from Donald Trump:



Regardless of how anyone feels about this president, I don't know that anyone thinks of him as being synonymous with environmental infrastructure and clean water. But taken at face value, it's easy to understand why our network predicted that this quote came from Clinton. The point here is that our network is shining a spotlight on *potential* areas to look for common ground similarities: it's still up to us draw our own conclusions from the results.

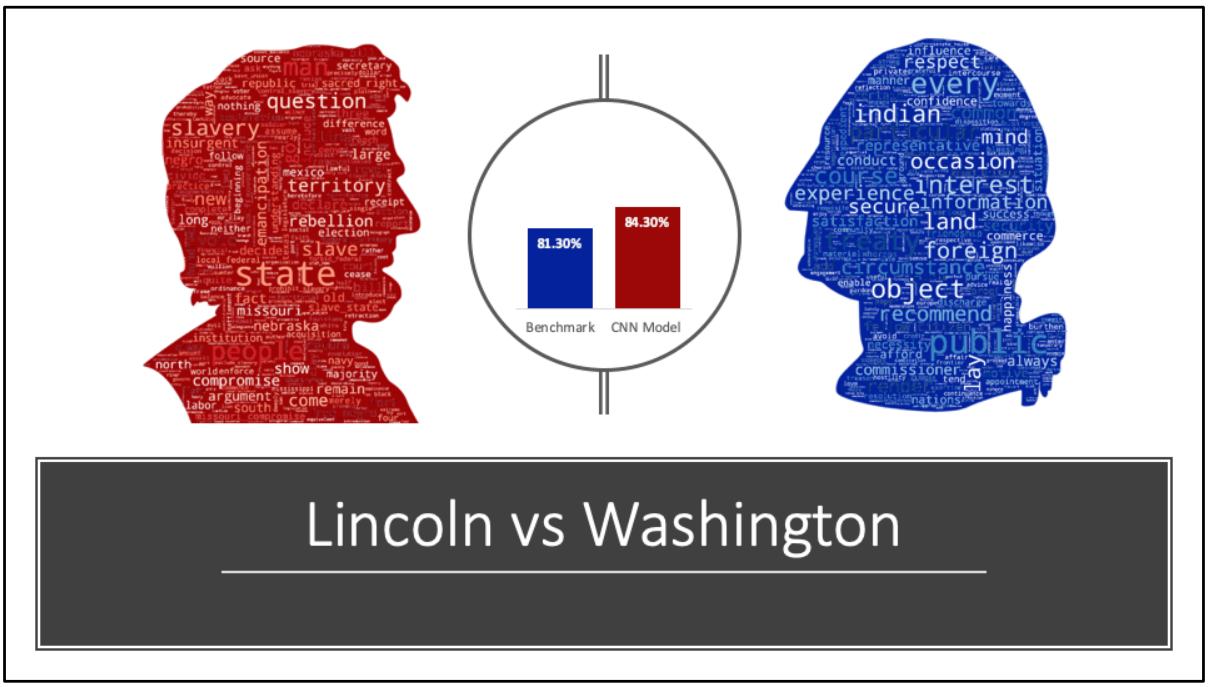
Unfortunately, we're not yet at a point where all the errors made are this easy to understand, so there's still more work to be done. As our networks are tuned and refined, and accuracy scores are driven even higher, we can give that spotlight more and more focus, thereby reducing our search area.

But before we get to that, let's apply this process to other presidential pairings and see what we get.

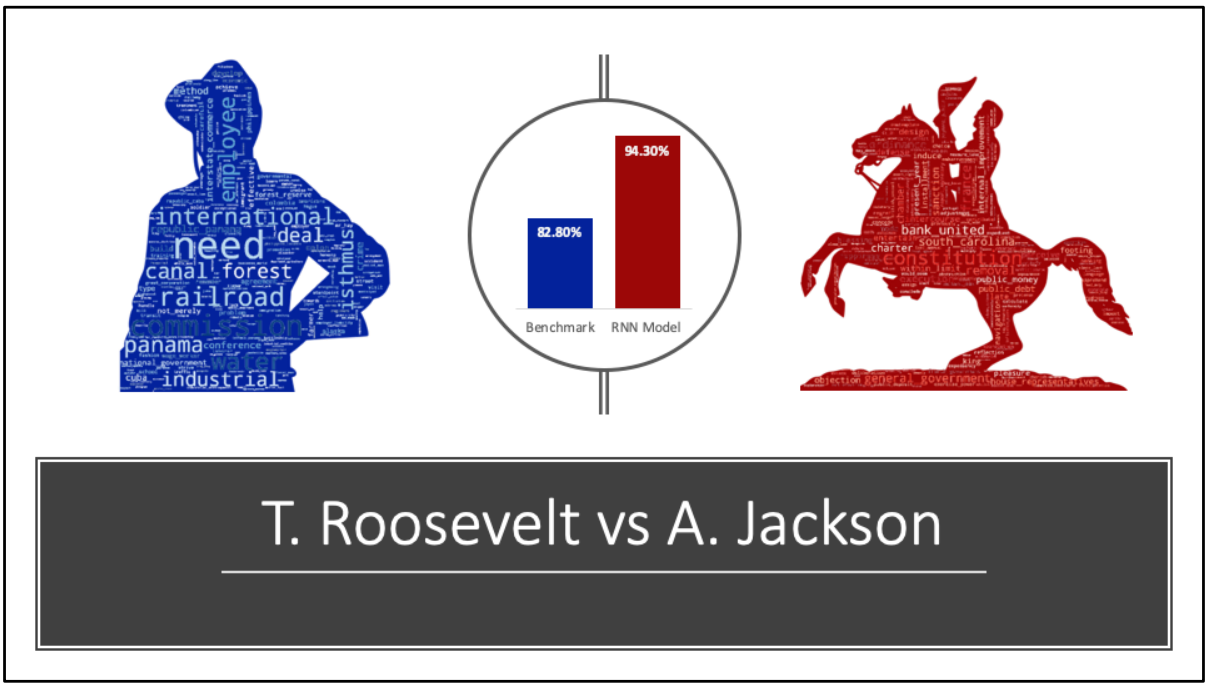


Starting with modern history, let's have a look at Barrack Obama and Ronald Reagan. We can see that traditional methods were able to make predictions with an accuracy rate of just under 75%, while a Recurrent Network was able to achieve a rate of 87.6%! This is also a significant improvement, but the lower score relative to Clinton and Trump suggests that these two may have a bit more in common than their modern counterparts.

If we have a look at their word usage, we can get a sense of the themes related to their respective presidencies. With Obama we see a big focus on health care, whereas with Reagan the focus is on the Soviet Union and Cold War.



Next, let's have a look at two of the biggest names in the nation's history: Lincoln and Washington. Here we see a Convolutional Neural Network produced the best results, though it only slightly outperformed traditional models. Of course, with Lincoln we see clear references to slavery, so no surprises there. Yet with Washington, identifying any clear themes is a bit more difficult. It leads one to suspect that the aim of his presidency, at least from the perspective of public persona, was to keep the country united, trying to not rock the boat too much while the country was in its infancy.



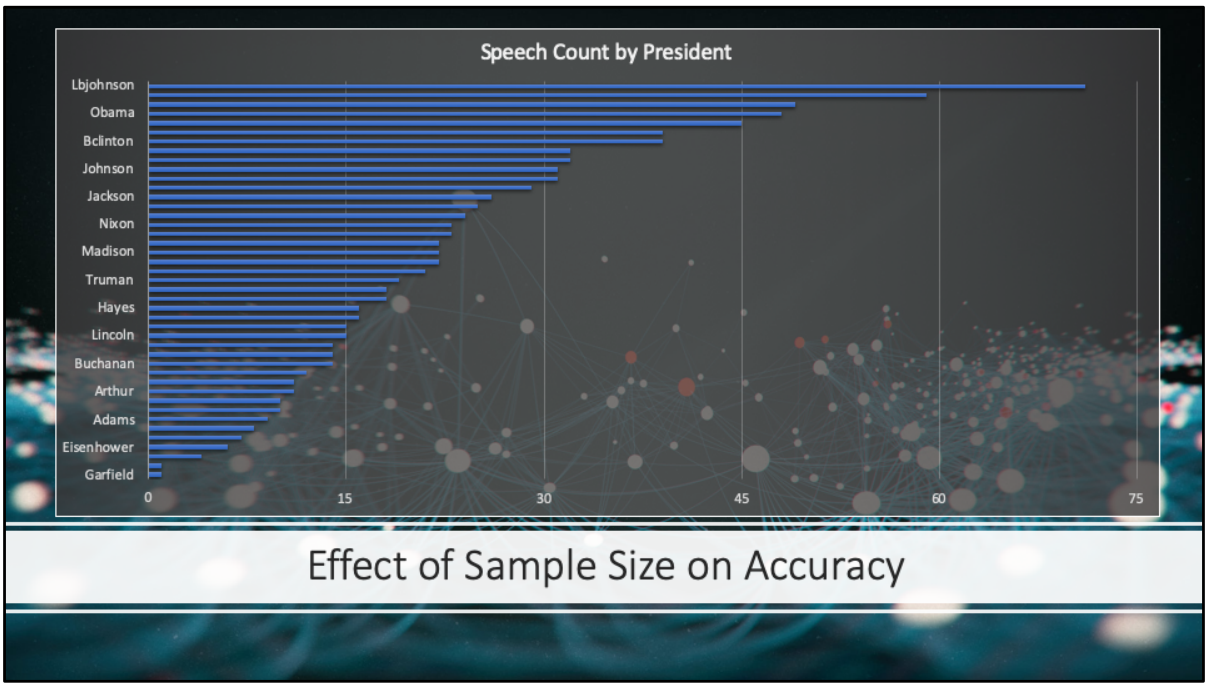
Let's have a look at one more: Teddy Roosevelt and Andrew Jackson. Here we find our highest performing neural network with a balanced accuracy score of 94.3%, and a great improvement over the benchmark.

And as we look at the differences in word usage, certain theme seems to be emerging: the greater the distance in time between presidencies, the more the word clouds seem to reflect differences in issues of the time than differences in the individuals.

Now, keep in mind this is just a first impression, and when we use these in conjunction with network predictions, we can start getting a more definitive picture.

So, where do we go from here? What can we do with these models? Well, the idea is that we can apply them to any two politicians or individuals in search of areas of common ground. In a country divided, I think it's good to look for opportunities to come together. Unfortunately, the model isn't quite ready for prime time and will need some refinement by improving both accuracy and efficiency.

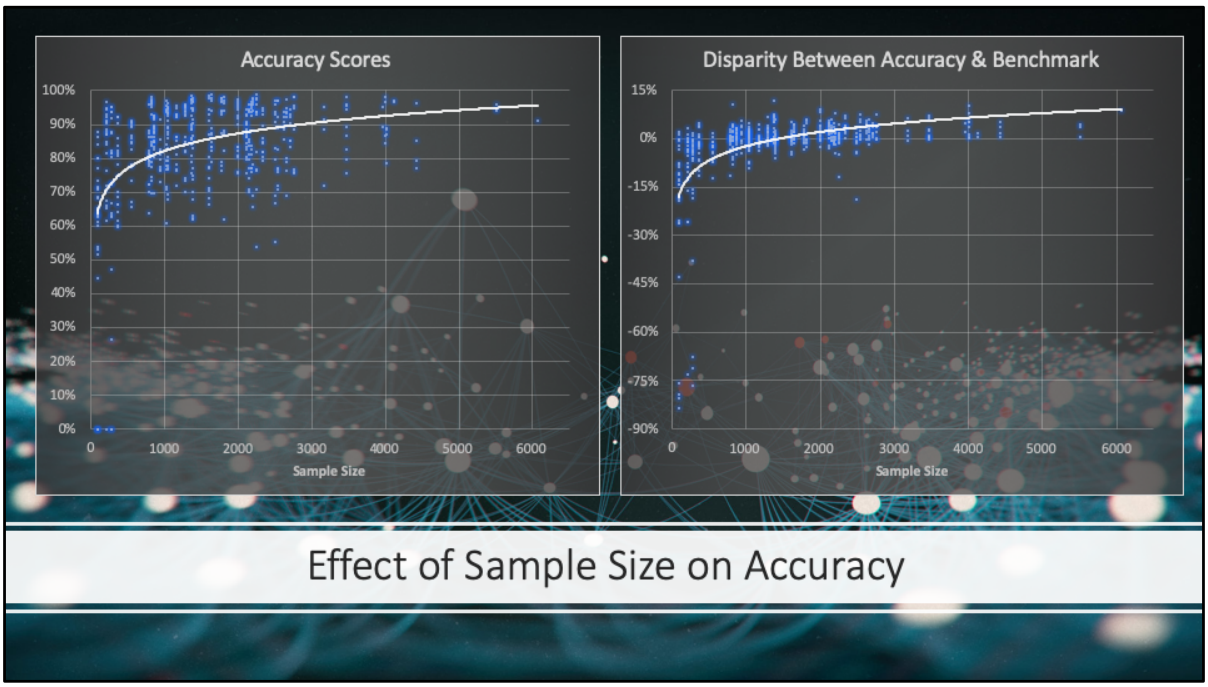
In order to do that, we need to gather more data so that we can get a clearer sense of which models and hyperparameters yield the most consistent results, as well as what type of data features are best suited for this type of analysis. So, I went ahead and applied randomly selected models to 500 randomly selected pairs of presidents and will share some of the initial findings here.



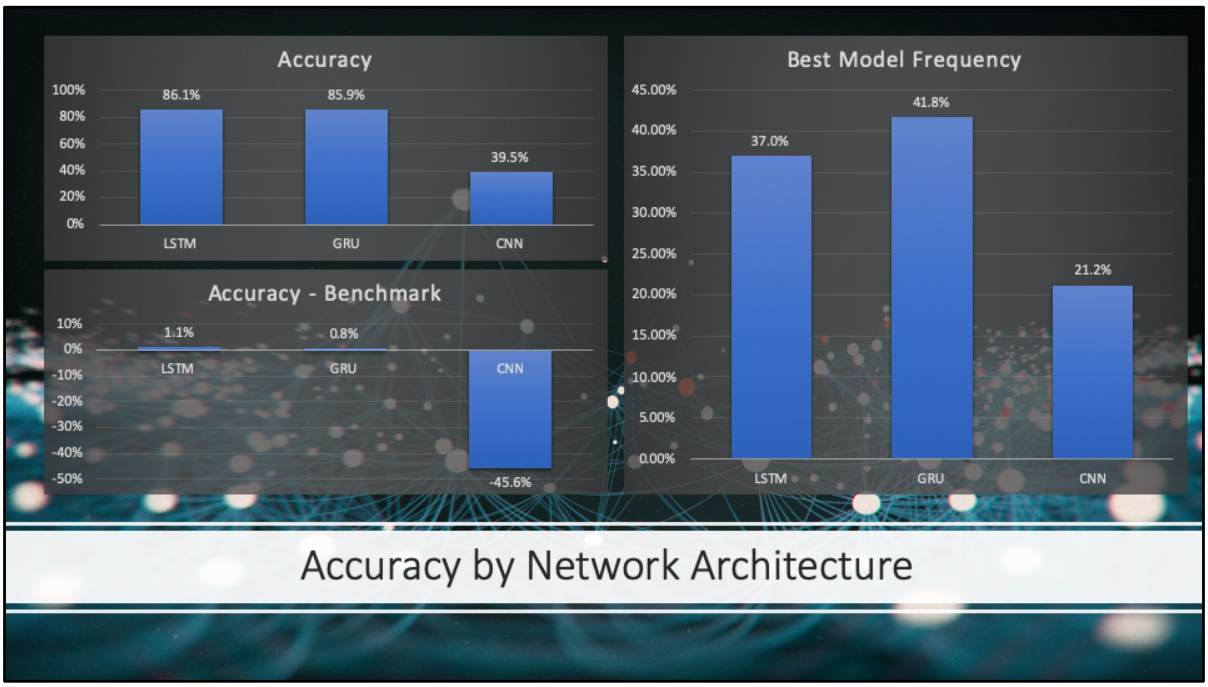
First off, one big issue is data imbalance. On one end of the spectrum we have Lyndon Johnson, who gave 71 speeches and for whom we have by far the most data, and on the other we have James Garfield who was assassinated after 100 days in office and whose only documented speech was his inaugural address.

If we wanted to perform a fair comparison of the two, we would have to select a sample of passages from Johnson that is equivalent in size to Garfield. As you can imagine, this is going to result in a great loss of data and creates a number of potential problems.

With that in mind, let's see how sample size impacts accuracy.



As we can see here, absolute network performance, as well as performance relative to the benchmark, greatly suffers when we have fewer than 500 samples. We see lower average scores and greater variation in the results, indicating that for such small samples, those results represent little more than random chance. This is something that needs to be taken into consideration when comparing other politicians in the future.

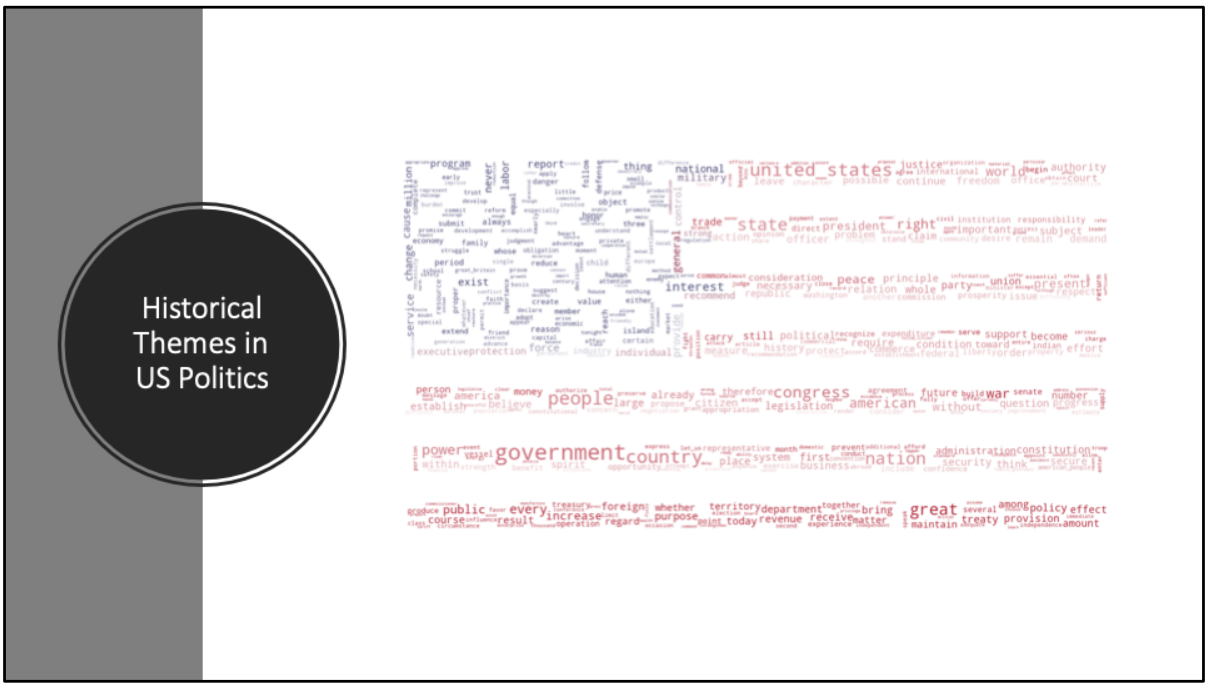


The other items we want to look at are network hyperparameters and architecture. As mentioned before, the two general types of models used were Recurrent and Convolutional networks. However, there were two different types of Recurrent networks used: one uses Long-Short-Term-Memory neurons (or LSTM), and the other uses Gated Recurrent Units (otherwise known as GRUs).

As we can see, LSTM and GRU models greatly outperformed Convolutional networks, varying only slightly from one another. The Convolutional networks generally performed rather horribly overall.

However, what is interesting is that these networks were still found to be the best performer in over 20% of the samples selected. This could either mean that there is a limited set of hyperparameters which make such models effective, or it could mean that these models perform better only when specific sets of data features are present. Either way, this is worth further exploration. And much work remains to be done, but the good thing is that there is a clear path forward.

From the observations thus far, the objectives of this project do seem plausible in principle, and we have a model building approach that will help us get there. Unfortunately, the current effectiveness of the approach still requires a fair amount of manual searching to find common themes, but tuning the model further, we can improve the accuracy, and thereby narrow the focus of our spotlight search.



So, with that, there is one last thing I'd like to leave you with. When gathering data for the 500 randomly selected presidents, error data continued to be tracked when the accuracy reached a sufficient level (defined as scores over 90% or when the network outperformed the benchmark). With this, one final word cloud could be created, which one might interpret as reflecting the common themes in US presidential politics throughout history.

Though more work will be required to figure out where we're going, it's always a good practice to reflect on where we've been.