Actuaries Institute Podcast
Title: Unpacking the YDAWG's 2020 US Presidential Election Twitter analysis

Date: 9 December 2020
Interviewer: Kriti Khullar
Speakers: Henry Ma, Ean Chan
Duration: 21:18

Kriti Khullar:	**Hello and welcome to the Actuaries Institute Podcast. My name is Kriti Khullar and I'm an Associate of the Actuaries Institute. Today I am joined by Henry Ma and Ean Chan who are both members of the Institute's Young Data Analytics Working Group, also known as the YDAWG.**
Kriti Khullar:	**In the three months leading up to the 2020 US Presidential Election, Henry, Ean, and the YDAWG conducted an analysis of the use of Twitter by both incumbent President Donald Trump, Democratic nominee Joe Biden, their respective parties and the public to extract meaningful insights on the campaigns and explore what interesting analytics can be done using the Twitter data. With the election now done and dusted, how accurate was their analysis? Let's find out! Welcome, Henry and Ean!**
Henry Ma:	Hello.
Ean Chan:	Hey Kriti, good to be here.
Kriti Khullar:	**Good to have you guys here. I'll go straight into it. So, talk us through the key points of the analysis.**
Henry Ma:	I can start talking about the first half of the analysis. So, we downloaded the tweets from several Twitter accounts we selected and clean those data first. The Twitter accounts we selected included, obviously the two candidates and their respective parties, plus some rather famous, or should I say infamous political personalities. And then with the clean dataset, we first did some descriptive analysis on the frequently used words and phrases, the candidates, and parties.
Henry Ma:	We presented the results in scaro tags and work cloud which looked pretty cool. The visualsation built, especially for scaro tags was quite useful because it allows users to search for keywords and see what the frequency is for the usage. And the next bit we thought would be interesting to do was to do a high-level

sentiment analysis on both the candidates own tweets and the tweets by the general public, who would directly commented or mentioned the two candidates and parties and see what's the reaction and the level of the positivity or negativity from those tweets.

Ean Chan: So, the next half of the article we had a bit of fun I guess compared to the descriptive analytics before. So we tried to generate some fake tweets based on the 10,000 tweets or so that we'd gotten from each candidate. And so, we'd actually fed that through a Markov chain and then produced tweets in the Twitter format. We changed some of the code behind the scenes to make them look like actual tweets, but always made sure to include fake tweet warning stamps on each of the tweets so that we didn't add to the fake news that's taking over Twitter at the moment.

Ean Chan: So, then we included other political personalities other than just the candidates. So, the Vice President, Mike Pence. The, well now, Vice President-elect, Kamala Harris. We had the GOP in there. I think we had Ben Shapiro as one of the fake tweeters. And then that was when we found that people had done a lot of analysis actually on Trump, big surprise.

Ean Chan: So, they were actually tools to create a fake Trump voice. And so, we were actually able to give our fake tweets his voice as well. So, it sounds a little robotic still, but definitely contains the essence of what Trump sounds like. Finally, we managed to do a Trump video to match the voice as well. And maybe egotistical of us, we got him to say some great things about YDAWG and our fake analytics group 'YDAWG Analytica'.

Kriti Khullar: **Being a member of the YDAWG group myself, I don't think it's egotistical at all. It's all fun and games. Thank you for that. Thank you for that Henry and Ean. So, in a nutshell, describe what type of research this analysis is categorised under.**

Ean Chan: So, to start off with as mentioned, it was mostly descriptive analysis, right? So it was like seeing the most common words used and there are ways of telling how much a word is used more by one candidate than the other and compared to the general population. And so you could see key favorite phrases and things like that. As Henry had mentioned, there was some sentiment analysis. And so we'd used hashtags of Donald Trump and Joe Biden, as well as just their names used in tweets. And sentiment analysis is a technique from natural language processing often called NLP. And that's a whole body of, I guess, interesting machine learning. For the fake tweets, we kept it relatively simple and we use Markov chains. So that really was just taking each tweet for a particular candidate and seeing what percentage of next words came off their any particular word.

Ean Chan: So, if the word ‘make’ was started in a Donald Trump tweet, we saw that some percentage of the time it was ‘America’ next often followed by ‘great’ and so on. And so that was the fake tweet section. So, for the Trump voice, we used a recurrent neural network. And so we use one in particular called WaveRNN. And so that’s a temporally sequenced neural network that uses the outputs from each stage as an input to the next stage of the neural network. So a bit different to a regular feed forward neural network but that was pretty good at producing a fake Trump voice, and there was a similar thing for the fake Trump video generation. I think that actually used the generative adversarial network as well in there. So, a bunch of machine learning bits and pieces all throughout but, people who are into machine learning love to share their stuff. So there's always things online to learn about and to use for your own purposes.

Kriti Khullar: **Definitely, definitely. Machine learning always is fun. So now, I just wanted to understand what was the inspiration behind putting this article together? How many people were involved?**

Henry Ma: Well, the short answer is we wanted to have some fun using analytics. I think everyone do some type of analytics at work but applying analytics to a very different context is definitely fun and then say interesting side project especially given we're working from home. So, we thought that's a really interesting exercise to explore. And then we look at what's going on in the world and obviously the US presidential election was one of the biggest things happening and there's so much hysteria both online and in all the other platforms about this election. And also, Twitter has become such a gold mine for data, especially sentiment analysis. So, we thought why not? So, people down did a very good work last year for the Australian e lection and apply that to the US context. And yeah, it turns out to be a very fun exercise and it was great to collaborate with different people in the group. I think we had about five, six people working in different capacity altogether and we completed working as pretty quick turnaround. So yeah, it was good fun.

Kriti Khullar: **Yeah. And good team effort as well. So, I wanted to also ask what are some of the high-level principles followed for this analysis?**

Ean Chan: So similar to what Henry said, we did do whatever sounded fun. In terms of the principles, we try to remain as unbiased as possible when doing any analysis. Trying to think about what areas we might fall into our own biases or pitfalls in analysis. For example, at one point we're doing the sentiment analysis and we'd included Trump as a key word as one of the words which indicated that it was a Trump tweet. But as you can imagine, the word Trump is

also just a normal English word and so it was also picking up a few different tweets which had nothing political about it at all. And so, we had to find ways around that such as only taking it when capitalized, or if there were other related things in there.

Ean Chan: And so, I guess part of that was some actuarial judgment type ideas where we did the first analysis and it wasn't really the results we were expecting. So we went back to see what could be wrong considering that one possibility is our initial expectations could have just been wrong. And so doing that at any point of the process was useful to producing the article we got to in the end.

Kriti Khullar: **I can see the professionalism of the actuaries coming in there as well, very good. So, I think Henry you were talking previously about the great work that was done previously. So how does the analysis of the 2020 US election compare to the previous election analysis from the 2019 Australian Federal Election that's been undertaken?**

Henry Ma: Yeah. It was a twin degree quite similar to what we did. The techniques were similar especially for the descriptive analysis part and then we both analysed some of the keywords and visualised in a scatter text which is quite interactive and meaningful. There are some differences compared to last year for what we did this year. So, for example, we added the part about the fake words and did a bit more work in that area. I think last year they used the open source Google API to do a bit more on the sentiment analysis. Yeah. So, I think they follow up quite similar principles, try to be neutral and well, try to have fun. So, yeah. It was a very good starter for the analysis we did this year.

Kriti Khullar: **I think just following on from that theme of doing similar analysis, what other areas can you see this analysis being applied to?**

Ean Chan: I think that a lot of these principles and ideas can be applied to almost any area. So for example, sentiment analysis is often used in reviews. So that's something we could do anywhere you find reviews. So, we could analyse Sydney restaurant reviews by cuisine and see what Sydneysiders like the most and plot that against what Melbourne, or Brisbane, have as their highest rated. And then we can compare maybe the actual ratings versus the kinds of words used. So maybe some people are more lenient about giving five stars and we can call out Brisbane for always giving five stars to restaurants that they might not believe as much in or something like that, right? We can extend that to LinkedIn comments if we want and get people's general sentiment on various actuarial employers. The possibilities are endless.

Ean Chan: I think Markov Chains in general are very good for impersonating other people, but also, I've seen articles written entirely by bots. So maybe our next Actuaries Digital article could be entirely based on every previous article and then a Markov Chain produced to just automatically generate one. Maybe we'll try and write in a few more sensible rules so it's not as chaotic but yeah. Maybe even a Markov Chain to produce my next appointed actuary report that I have been put into. Who knows? But there's basically a lot of different data sources out there and all of them can be analysed in some way and I guess that's what we're trying to do a bit at YDAWG making people realise that you can analyze anything and it's all ready, and it's all out there for us to analyse.

Kriti Khullar: **Definitely, definitely. And just touching on the chaotic comment you made, we'll just put some rules around it. So, what were some of the challenges while going through the project and how were they resolved?**

Henry Ma: There are a couple of challenges, but I'd also say it's part of the fun. So, one of them was we got different people working on the same document at the same time. That takes a bit of coordination and planning to divide a task and in terms of keeping us honest in terms of when to complete what sections and so on. And also, we had a bit of turnarounds, I think we had about three to four weeks before the election because we need to complete that by then and also allow some time for the Institute to publish that. So, yeah, a bit of a turnaround but it's nothing we're not familiar with as a consultant. So that's all good. Yeah. Now that's some of the challenges

Kriti Khullar: **I think, yeah. Tight timelines are probably sometime the flavor of the COVID year as well, so you just extend it here. But there were hopefully some favorites as well. What were some of the favorite parts of this project?**

Henry Ma: Yeah. For me it's definitely applying analytical techniques to non-financial services context. Because mainly why I focus on is this financial services banking and so on. Actually, the first time I did some analytics that's substantial enough to call it analytics and being able to apply that to a very different context is very fun and refreshing.

Ean Chan: I agree with Henry. I always do enjoy applying things, applying actuarial and analytics ideas to places I usually don't. I think for me over this project, I really enjoyed each time we came up with a new idea, right? Because we'd started off with trying to mostly replicate the 2019 article but working agile meant that we could come up with an idea and then whoever was free and ready could work on it to try and produce something good out of it. So I did enjoy each of those moments when someone was like, "We could

do this." And I was like, "Why don't we do that?" So that was good for me.

Kriti Khullar: **That's lovely. So, speaking of bits that you were able to put in, what were some bits in the articles that you wanted to put in but weren't able to?**

Ean Chan: So, with the fake tweets, I actually did this, but we couldn't put it into the article. I'd made a quiz where you had some of the fake tweets and some of the real tweets for each person mixed in together and you had to distinguish which was which. Which was pretty fun so I had shared that with some colleagues and friends and some of them were actually pretty hard to tell that they were fake. I will say that some Twitter accounts lend themselves to being impersonated more easily than others. Another thing that we thought about putting in was something to do with fake tweet bot detection, right? It was very closely linked to some of the stuff we were doing and there was, as we talked about earlier, a lot of fake news out there. And there are pretty, I guess, sophisticated ways to try and catch them out but we hadn't gotten around to doing anything ourselves with it. But if anyone wants to know more about it, there's plenty online where people talk about how they catch these kinds of bots.

Kriti Khullar: **That sounds all very interesting. I know I'll be looking up as well. So, what other direction would you have taken on this article?**

Henry Ma: Yeah. For me, very obvious the thing is moving away from just analysing Twitter data to doing some symbol predictive. For example, predicting the primary results, or even the ultimate Presidential Election results. That if you spend some time searching and Googling online and you see there are tons of models out there ranging from pretty basic and high level to very sophisticated models. So, I think we felt like with the time we were given and also the level of commitment that required to learn and investigate what is appropriate to use, it's quite a non-trivial task. So, we didn't go into that direction but that could have been something we did.

Ean Chan: Similar to that, there were some other election related ideas but not Twitter related ideas that we had including doing analysis on turnout by state for the election, all that public daughter as well. So you can actually find some interesting correlations between the turnout for a state in the 2016 election and whatever you like. So average temperature of the state, or percent unemployed in the state in 2016. And then you can either just use that as the fun little tidbit it is, or you can use that to also try and predict 2020

turnouts. But being the unprecedented year that it is, I think that might've fallen a bit flat. There was plenty of data circulating the election that we thought about doing analysis on.

Kriti Khullar: **So, I hear basically there's this lots of data available which is definitely one of the resources, but what was some of the, I guess, resources you drew upon to deliver this project?**

Henry Ma: Last year's article and GitHub.

Ean Chan: Twitter. We coded in Python. So there's quite a few Python packages out there. The main ones like Pandas, we use a Markov Chain one, there was a Scatter Text and Word Cloud one. But also we used Google Colab. So, in the same way Google Documents allows people to all edit the same word document online, there's actually, well we used the Python one, a Python notebook online. So similar to a Google doc, this is just a Python notebook run by Google that everyone can access at the same time. And so we could all edit that for when we wanted to make changes to the code. So I would recommend that, that's a very useful one. In terms of, I guess learning skills, for me YouTube is a big one to be completely honest and also Kaggle, lots of resources on Kaggle, both data sets and explanatory notebooks to guide you through the process.

Kriti Khullar: **Thank you for sharing those tips. I think that's probably quite useful for some of our listeners who're wanting to move away from the sort of analysis into some predictive analysis and even, I guess, trying to create such an article. But what are some upcoming projects that you are looking forward to?**

Henry Ma: We're still planning out what we're to do next and I'd really welcome any suggestions or feedback from the audience of the podcast on what topic, interesting stuff they want to see. But for now, I reckon a that stay tuned.

Kriti Khullar: **Thank you for that. So that's all we have time for today. Thanks for joining us Henry and Ean.**

Henry Ma: Thank you.

Ean Chan: It’s been a pleasure.

Kriti Khullar: **Listeners we hope you enjoyed this discussion. To listen to more Actuaries Institute Podcast, please visit the podcast section of Actuaries Digital. As always writing with your questions and comments on the show, we'd really appreciate hearing from you. I'm Kriti Khullar, bye for now.**