# Slide 1

Hi, My name is Gordon Hew and I’ll be presenting the AI Incident Database Case

# Slide 2

The presentation will take on the following format.

First, we’ll go over the background of the AI Incident Database Case

Next, We’ll cover the Objective and then the Data & Approach.

After that - We’ll then go over the Analysis of the data and wrap up with my Conclusions

# Slide 3

So first – what are AI Incidents? Also why are they important?

According to the Artificial Intelligence Incident Database (AIID):

*“AI incidents are events or occurrences in real life that caused or had the potential to cause physical, financial, or emotional harm to people, animals or the environment.”*

AI incidents are important because they can have unintentional and unwelcome real-world consequences

This includes problems in Injustice / Inequality where systems can create racial bias in fields such as healthcare and probation systems leading to asymmetrically poor outcomes for a given group of people.

Economic Loss can be the result of trading algorithms can behave badly and lead to market crashes

Death and Injury has occurred where intelligent systems such as those found on the Boeing 737 Max can lead to fatal consequences

# Slide 4

So this leads us to the objective – what do we want to do with this AI Incident data.

We want to apply natural language learning techniques to investigate trends in AI incidents, dig into any interesting findings, and raise areas of concerns around AI

# Slide 5

The analyzed data is a course provided sample of data extracted from the AI Incident Database.

This extraction involves a time period from 1983 to 2020 and contains 92 AI Incidents.

These 92 AI incidents contain 1225 reports. This means that one incident can have more than one report.

Reports consist of a ID that maps it to a unique incident, the title of the report, the source of the story, and the body of the text.

Our approach to analyzing this data is to first figure out how we want to divide up the data for processing.

I’ve divided the time periods for the text in three periods.

The first period is incidents before 2011 – this is before deep learning and big data really took off. Furthermore, there’s fewer incidents and also years within this spand that didn’t have any incidents.

The next two periods is an even division of the remainder with each period being five years.

The hope of dividing the incidents in these large periods is to be able to track large general trends in terms of topics.

The next step is to clean and prepare the data for processing.

This means that we remove stop words, URLS, remove brackets, replace abbreviations.

This text data is then vectorized in preparation for being fed into our cluster models.

In my analysis, I’ll be using two unsupervised learning methods for topic cluster analysis. The first will be LDA which scans the words in the documents and assigns probabilities that a particular word in a document belongs to a topic. The second is K-means topic clustering which uses the process of assigning vectorized text data to the nearest clustered mean. This approach is then combined with emotional sentiment analysis.

# Slide 6

Before we get into clustering we want to get a general sense of the frequency of the AI incidents and reports over time.

We see in the first chart there’s only a handful of incidents if any from 1983 to roughly 2013. From there the number of incidents picks up.

In the average reports per AI incident, we actually see that there were typically many reports for a single incident until about 2013. Furthermore, the number of total incident reports remained fairly low until it spiked in the mid 2010s which corresponds to the wide-spread use of big data and deep learning.

# Slide 7

This word cloud gives us an idea of what news sources were submitted to the AI Incidents databases.

There appears to be a large British bent with the guardian, dailymail, bbc, telegraph, the register, and the independent as larger sources of news. Otherwise, we see a number of American mainstream news outlets such as CNN, USAToday and the NyTimes and also some other more focused outlets such as cnet, engadget, and popular mechanics.

# Slide 8

Looking at our first time period, we specify 5 clusters in our model but we are only going to examine the top 3 for expediency purposes. We see the five clusters on the right and 2 sets of clusters have some overlap. On the left, we see the top 30 most salient terms which is a metric that tell us what are the most useful words in identifying a topic along with the term frequency. Among the top 4 terms, we see crash, amazon, flash, and petrov – and we’ll see how salient these are when we drill into the topics.

# Slide 9

This LDA visualization tells us that one big topic area during this time period was when Petrov, a Soviet lieutenant colonel decided not to retaliate against the US when his system told him that five missiles had been launched from the US thus preventing nuclear war.

# Slide 10

This second topic tells us that there was a flash market crash during this time period caused by Navinder Sarao who spoofed orders by creating artificial demand which resulted in automated trading systems to mount loses.

# Slide 11

This third topic tells us that Amazon had an AI incident that removed the sales rank from gay/lesbian themed books because a model had classified them as Adult products.

# Slide 12

Lastly, the k-means clustering yielded similar clusters based on the terms on the left. Sentiment from these incidents seem to be both positive and negative. With a lot of negative sentiment relvolving around fear and the positive sentiment mostly in trust.

We see that in this time period that AI incidents are dominated by rather specific headlines rather than a general theme.

# Slide 13

Our second time period appears to have more companies and products among the most salient terms. Let’s examine the first topic.

# Slide 14

The first topic cluster appears to be more general than previous topic clusters we’ve investigated. Diving deeper in the salient terms on the right and searching a few key terms, we discovered that Professor Sweeney at Harvard found that Google searches and results had a discriminative bias.

# Slide 15

The second topic cluster for this time period resolves around Ibrahim Diallo who was fired by an automated system that step-by-step locked him out of his workplace when his manager forgot to renew his contract.

# Slide 16

The third topic cluster for this time period also appears to be less specific at first glance as it revolves around amazon, alexa, and echo. However, the dollhouse term refers to an incident where a child made an order for a dollhouse through an echo.

# Slide 17

The k-means yield similar topic clusters and there is still that mix of negative and positive sentiment. However, there appears to be more anger in this period – possibly due to the book removals and the accidental firing.

# Slide 18

In our final period, we see that most of the salient terms seem to revolve around self-driving cars and facial recognition.

# Slide 19

Our first topic cluster based on these terms revolves around a self driving Tesla car accident.

# Slide 20

Our second topic cluster appears to be more general and it’s more difficult to derive what the context of the documents in this cluster as the terms appear to be more disparate and none of them alone are powerful indicators of signifying belonging to this topic cluster.

# Slide 21

Our third topic cluster revolves around the use of facial recognition across applications and countries as there has been incidents of widespread adoption in China to use facial recognition to identify and apprehend criminals.

# Slide 22

In this period we see still see mixed negative and positive sentiment across the board. There’s also less anger in this period when compared to the previous period.

# Slide 23

So using topic clustering across different time periods, we found that in the first period, the reports were mostly about nuclear war, the stock market flash crash, and Amazon sales rank.

The second time period revolved around discrimination by google search, automated work firing, and Amazon Echo.

The final time period revolves around self driving cars and facial recognition and surveillance.

We see that the AI incidents are moving from the more abstract – large scale events to becoming interwoven to everyday life such as your work and driving.

Sentiment also appears to be a mixture of positive and negative where the negative emotions are mostly around fear and anger.

From this analysis we see that AI incidents has impacted society from where it didn’t impact us at a specific individual level to now being part of our everyday life. I think this analysis raises the point that we need to look forward in terms of deciding what we will allow AI to augment / perform for us on our behalf and what we won’t.

Thank you for your time and I hope you enjoyed this presentation.