# 

# **Milestone 9: Final Capstone proposal**

# **Deep Fake Detection and the Implications of Deep Fakes on Democratic Processes**

# 

Asmaa Aly

CP192 -Levitt, Th@18:00 Buenos Aires (Spring 2020)

Computer Science Major: Data Science and Statistics

Social Sciences Minor: Government and Politics

# Tables of Contents

[Author’s note](#_3dy6vkm) 2

[External Validation and Critique](#_1t3h5sf) 2

[Stage 1: Proposal](#_4d34og8) 2

[Stage 2: Project](#_2s8eyo1) 3

[Abstract](#_17dp8vu) 3

[Background and context](#_3rdcrjn) 4

[Similarities and Differences](#_26in1rg) 5

[Challenges](#_lnxbz9) 7

[Description of the proposed project](#_35nkun2) 7

[References](#_44sinio) 9

[Appendix A: Goals, specific objectives, key results or deliverables, and deadlines pertinent to your project.](#_2jxsxqh) 11

[Appendix B: HCs and LOs](#_z337ya) 12

[Appendix C: Potential sources of relevant support, advice, or feedback](#_3j2qqm3) 17

[Appendix D: Prototype](#_laf0q4gthxzb) 18

[Appendix E: HCs and LOs applied in this proposal](#_1y810tw) 19

# **Author’s note**

New project or continuation: This project is the continuation of my previous assignments in the spring. I did change my project from an online mentorship program to the current proposal of a comparative analysis paper that discusses the state of art of the deep fake detection techniques.

One project or two: This PDF has only one project idea with no external dependency.

# External Validation and Critique

## Stage 1: Proposal

For stage 1, I consulted Professor Sterne and my previous internship mentor. Both assured that the project is feasible under the time constraint. Professor Sterne also highlighted how the depth of the application of LOs from my major is a key strength of the proposed idea.

## Stage 2: Project

For the project, I am consulting my mentor in my previous internship. He got his Ph.D. in computer vision applications from MIT. He gave me some feedback on the literature review that I submitted before, and he guided me to where to find other papers that can be a representative sample of the state of art of deep fake detection. I am also planning to grow my network during the summer to have another mentor whom I could consult during the execution phase of my project.

I added a third section of the academic paper: a policy recommendation memo for social media companies, based on the findings of the paper.

# Abstract

Deep Fakes are falsified videos generated by Generative Adversarial Networks (GANs). Apps like Face Swab can easily produce a counterfeit of original videos and photos by using generative adversarial networks. The technology was used to create fake videos of targets increasing the chances of spreading fake news on the internet. My capstone aims to use insights from deepfakes-detection-state-of-the-art papers to propose an alternative model and policy recommendation for social media companies. For my deliverables, first, I am aiming to develop a comparative analysis by replicating the five most cited papers. Second, I am aiming to develop an alternative approach using the insights from the comparative study, optimizing deepfakes detection and a benchmark dataset. Third, a decision memo for social media companies to adopt deep fakes detection for each video before it is uploaded.

# Background and context

With the increased accessibility of video manipulating and video generating applications like Face2Face and DeepFake (Harwell, 2018), the spread of deep fake videos, as a tool to spread fake news and allegations, is a threat to human rights around the globe (Citron, 2017). Despite their dangerous potentials, cyber criminals are posting videos of target figures to cause harm, instability, and destruction. Deep fakes have a destructive potential as social media enables anonymous posting of these videos with no regulations. In 2019, Facebook announced that they are developing tools to detect deep fake videos after they are posted to remove them (Schroepfer, 2019). The social media giant announced a Kaggle competition in December 2019, along with Microsoft, to develop an algorithm that can detect these videos (Deepfake Detection Challenge, 2019).

Deep fake videos are generated by continuous interpolation of the target face to create fraudulent videos using these views in two different ways (Guera & Delp, 2018). First, the interpolation of the target can be used to create another video of the target where the facial reactions and words are changed. Second, the interpolation can be used to swap one target's face with another one creating an almost realistic video that can tarnish a target's reputation.

Some of the common deep fake tools are free and can be used anonymously which makes the powerful technology a handy tool for cyberbullying or the spread of fake news. For instance, Rana Ayyub, an Indian author and an activist, had a deep fake video of her which her face was used on a pornstar’s body. The video was shared thousands of times leading to death threats and cyberbullying (Harwell, 2018). As a result, MS. Ayyub had to delete some of her social media accounts for her safety. This case, along with others, suggest that the only way to control the spread of deepfakes is to identify them before they are uploaded online, creating an automated filter that can prevent a cybercrime from happening on a certain platform (Citron, 2017).

The state of art of deep fake detection is growing. A numerous amount of papers discussed different methods and techniques that the researchers developed to detect these videos on social media, mainly Facebook and Twitter. In my literature review, I am exploring the question of what the differences, the similarities, and the challenges that each of these methods have.

I will examine Deepfake Video Detection through Optical Flow based CNN (Amerini, Galteri, Caldelli & Bimbo, 2019), Synthetic Portrait Videos using Biological Signals (Ciftci, Demir & Yin, 2019), Exposing DeepFake Videos By Detecting Face Warping Artifacts (Li & Lyu, 2019), Compact Facial Video Forgery Detection Network (Afchar et al., 2018), and Exposing Deep Fakes Using Inconsistent Head Poses (Yang et al., 2018).

## Similarities and Differences

All five research projects break down the videos into multiple video frames. Another key similarity here is using reverse engineering as a framework to develop their approach. For instance, reverse engineering applications that use generative adversarial networks highlighted the key processes that made each video. Nirkin, Keller, and Hassner derived a recurrent neural network (RNN) algorithm that can swap faces and adjust for the pose and the variations in the facial expressions. Their approach uses continuous interpolation of the face views and a face blending network to preserve the target skin color and lighting conditions. The blending network uses Poisson optimization for this goal as well.

Each paper detects some of the errors that can result at each stage of making a deep fake video. Amerini, Galteri, Caldelli, and Bimbo’s approach compares the similarities between the different video frames. If the similarities exceed a certain threshold, then the videos are real. Ciftci and Demir use spatial features of the faces as biological signals to determine whether they are consistent across different video frames. Lyu and Li use the resolution of the different images in the video frame to determine whether the videos are fake or not. The premise of this paper is that convolutional neural networks can only generate video frames with a limited resolution. Thus, if the quality of the video frames is relatively low and inconsistent, the videos are probably fake. Afchar, Nozick, Yamagishi, and Echizen use the same premise of the quality of the videos in the different video frames in each video. Their approach is to use a low number of layers in their networks to detect the slight differences in the mesoscopic properties of the images. Twas able to detect deep fake videos generated from applications like DeepFake and Face2Face with 98% and 95% accuracy rates respectively. Lastly, Yang, Liu, and Li try to compare the coordinates of the head poses in each of the video frames of the video and compare the observed coordinates of the head poses in the videos with a 3D prediction of how these coordinates shall be in real videos and real images.

## Challenges

The datasets used for each paper are different. Thus, comparing the key metrics reported for each paper is hardly possible. For instance, Lin and Lyu use deep fake videos found on the internet as their test set whereas Afchar, Nozick, Yamagishi, and Echizen use videos generated by Face2Face and DeepFake applications. Moreover, there is lack of a comparative analysis or literature review in each paper. Instead of comparing how each method compares to the other, the comparison is based solely on the mere fact of the nuanced approach proposed, without taking into consideration how each approach adds to the literature.

# Description of the proposed project

My intended approach is to compare the state of art papers that present different methods of detecting deep fake videos. I picked the most cited papers, proposing high-accuracy detection techniques. All of these papers shared the data that they used to generate their results, so I am aiming to replicate these results. Moreover, since all of the papers are using different datasets to train and test their models, my replication is going to use the dataset made by Facebook and Microsoft as a way to establish a benchmark for a fair comparison of each method. In case I progress on my project, I will use the findings of each of the approaches to develop another technique using deep learning to detect the fake videos. Lastly, based on the findings of the paper, I will develop a policy memo for social media companies.

My internet intended deliverables are a comparative analysis paper that compares the different methods of deep fake detection using the Facebook and Microsoft dataset posted on Kaggle. In addition, the paper shall propose another technique based on the findings of the comparative analysis provided. There are two reasons why I am aiming to have these outcomes. First, I am planning to go to graduate school after Minerva, and after consulting with my capstone professor in the fall, I realised that working on a paper that can be published would add a significant value to my application. Second, there is a lack of papers that have a comparative framework of how each method works (Engler, 2019). The fact that deep fakes are relatively new makes my paper a new addition to the growing field.

So far, I was able to replicate the Ciftci, Demir and Yin’s using the video data on the Kaggle as shown in Appendix D. Finding the replication code for each paper is a challenge. That’s why I have two plans for my deliverables. The first one is just the replication of the methods on the dataset I found. In case I don’t find the code, I will produce my own aiming to replicate the same method. The second one, best case scenario, includes the first plan in addition to another proposed method that I will develop based on the findings of my comparative analysis.

The key metrics of the project are training time, validation score, training score, recall, precision, mean square error, mean absolute error, and model training loss– taking into consideration the constraint of low-resolution videos, where the detection of facial expressions and views is harder as the image quality does not allow for these small details to be picked up by the algorithms.

# References

Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). MesoNet: a Compact Facial Video Forgery Detection Network.

Amerini, I., Galteri, L., Caldelli, R., & Del Bimbo, A. (2019). Deepfake Video Detection through Optical Flow Based CNN. 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 1205–1207.<https://doi.org/10.1109/ICCVW.2019.00152>

Ciftci, U. A., & Demir, I. (2019). FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals. ArXiv:1901.02212 [Cs].<http://arxiv.org/abs/1901.02212>

Citron, D. (2017, July). How deepfakes undermine truth and threaten democracy.<https://www.ted.com/talks/danielle_citron_how_deepfakes_undermine_truth_and_threaten_democracy/transcript>

Deepfake Detection Challenge. (2019, September 11).<https://kaggle.com/c/deepfake-detection-challenge>

Engler, A. (2019, November 14). Fighting deepfakes when detection fails. Brookings.<https://www.brookings.edu/research/fighting-deepfakes-when-detection-fails/>

Guera, D., & Delp, E. J. (2018). Deepfake Video Detection Using Recurrent Neural Networks. 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 1–6.<https://doi.org/10.1109/AVSS.2018.8639163>

Harwell, D. (2018, December 30). Fake-porn videos are being weaponized to harass and humiliate women: ‘Everybody is a potential target’—The Washington Post.<https://www.washingtonpost.com/technology/2018/12/30/fake-porn-videos-are-being-weaponized-harass-humiliate-women-everybody-is-potential-target/>

Li, Y., & Lyu, S. (2019). Exposing DeepFake Videos By Detecting Face Warping Artifacts. ArXiv:1811.00656 [Cs].<http://arxiv.org/abs/1811.00656>

Nirkin, Y., Keller, Y., & Hassner, T. (2019). FSGAN: Subject Agnostic Face Swapping and Reenactment. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 7183–7192.<https://doi.org/10.1109/ICCV.2019.00728>

Schroepfer, M. (2019, September 5). Creating a data set and a challenge for deepfakes. Facebook AI.<https://ai.facebook.com/blog/deepfake-detection-challenge/>

Yang, X., Li, Y., & Lyu, S. (2018). Exposing Deep Fakes Using Inconsistent Head Poses. ArXiv:1811.00661 [Cs].<http://arxiv.org/abs/1811.00661>

# Appendix A: Goals, specific objectives, key results or deliverables, and deadlines pertinent to your project.

Figure 1. A screenshot of my notion page for the capstone. Here is a full link for the board exported as PDF. [Link](https://drive.google.com/file/d/1xen6DqJMy4jflaDlB6Lu8elSKGe8GLrf/view?usp=sharing).

# Appendix B: HCs and LOs

Specify 10-25 HCs that will be relevant to your proposed project in addition to the set of 5 HCs that will be assessed for all final Capstone work products (#professionalism, #organization, #evidencebased, #sourcequality, #responsibility

Table 1. HCs from cornerstone courses. FA stands for formal analysis, MC stands for multimodal communication, and EA stands for empirical analysis.

|  |  |  |
| --- | --- | --- |
| HC | Course | Application |
| #studyreplication | EA | In my paper, I will be reproducing the neural network architecture using the benchmark dataset. The replication aims to highlight   1. The validity of the results reported by the authors 2. A robust comparative analysis of the different methods using the following key metrics: accuracy, recall, precision, mean squared error, loss rate, training time, and overfitting. |
| #probability | FA | Classification using neural networks relies on different activation functions that transform the outputs of the neurons to probability. For example, the output layer in my neural networks determine whether a video is classified as fake or original using the outer layer activation function: softmax or sigmoid. Both of these functions generate probability of data points to be in one class or the other. In the paper, I will aim to explain the probability distribution that results from each of these functions. |
| #dataviz | EA | Since I am working with digital imagery, dataviz is a must to communicate the results of my analysis. For example, in addition to printing the scores of precision and recall– my metrics– I will be using the Receiver Operating Characteristic (ROC) plot to communicate the same key metrics. I will also be using loss function and accuracy graphs over the number of steps or epochs when I am training the neural networks. Finally, I will be using the HC to plot the data throughout my analysis to showcase the pre-processing stages followed. |
| #rightproblem | EA | The spread of Deep Fake is a complex problem that involves multiple actors and variables, i.e. lack of law enforcement and the spread of free GAN-powered applications for public use. Instead of focusing on each aspect, I decided to focus on the root problem: detection to limit the spread of the videos from the first place. |
| #breakitdown | EA | Classifying a whole video is computationally expensive. To overcome this obstacle in my prototype, I analyzed the anomalies in the videos by segmenting the video to multiple video frames. These frames are then used for detection. This approach will be adopted for the final paper as a pre-processing stage, where the number of the video frames segmented is a hyperparameter. |
| #gapanaylsis | EA | In my comparative analysis, I am including the different approaches each paper proposes. I am evaluating the strengths and the weaknesses of them based on the key metrics, i.e. training time and overfitting. Based on this analysis, I am proposing another approach that tackles the weakness of the previous algorithms. |
| #context | MC | There are two applications for this HC. First, in the Background and Context review illustrates the context and the motivation behind the pursuit of this project.  Second, the interpretation of the key metrics are subjective based on the context of the problem. For example, if one training algorithm takes substantially longer to train and evaluate with a high value of accuracy, it would be constrained by the computational resources. Thus, a small tech startup won’t adopt it for instance. |
| #constraints | EA | There are three applications for this HC. First, since my internship became remote, I am unable to use the High Performance Computing resource to build and train the neural networks I am reproducing. Given this constraint, I am upgrading my account on Google Colan to increase the GPU speed.The second constraint is the time. In Appendix A, I managed to form a solid plan that has important deadlines and milestones that are achievable in the near future.  The third constraint is the absence of reproducible codes from the papers that I chose. Thus, I have to implement the neural network architecture solely based on what the authors highlighted in their methodology. |
| #optimization | FA | Optimization is a broad term. Training the models and tuning the parameters to achieve the highest accuracy without overlooking overfitting is an optimization problem. I am planning to explain the mathematical expression of how the neural network updates the weights/parameters in each iteration/epoch. |
| #variables | FA | The code for the prototype and the final product have numerical, categorical, intermediate, dependent, and independent variables. For each function and code section, there will be comments that describe the purpose of the variables, the type of it, the type of attributes it holds– i.e. string or int. |
| #critique | MC | Based on comparative analysis in my paper, the analysis of the key metrics is essentially a critique of the approaches based on evidence. |
| #audience | MC | The level of detail and sophistication of the resultant paper is geared towards experts in Generative Adversarial Networks and Convolutional Neural Network researchers. Thus, adapting to the guidelines in the research community is a key metric for my paper. Here is a [reference](http://iccv2019.thecvf.com/submission/main_conference/reviewer_guidelines) to some of these guidelines. |
| #algorithms | FA | Machine learning models are based on mathematical models that get trained on data. The process of training the model is an algorithm itself; it involves a problem and a series of steps followed to solve the problem. The optimization of these algorithms/parameter tuning uses a series of steps. For example, the CNN has to pass the data to the input later then to an activation function.. outer layer. I am planning to use diagrams to show the architecture of the neural networks. Moreover, adding comments and implementing the mathematical models proposed as a code requires certain steps, i.e. pre-processing |
| #designthinking | MC | Over the course of last year, the initial idea of proposing a new machine learning model that classifies deep fakes took an iterative process. The initial goal was the creation of a novel solution, but now, it is a comparative analysis of the state-of-art papers. The change is the result of incorporating the feedback that I got from the assignments and the milestones. |

Specify 5-10 LOs from your major(s). Students with two majors must include at least 2 LOs from each college.

Table 2. LOs from my major.

|  |  |  |
| --- | --- | --- |
| LOs | Class | Use |
| #neuralnetworks | CS156: Machine Learning for Science and Profit | Deep Fakes are created using Generative Adversarial Neural Networks. The literature review discusses how GANS are implemented. The literature review also discusses how neural networks can be used to extract features of video segments. As expected, the state-of-art papers propose different implementation of deep fake detection using neural networks. |
| #modelmetrices | CS156: Machine Learning for Science and Profit | The key metrics in my comparative analysis are based on what I learnt from this LO: training time, validation score, training score, recall, precision, mean square error, mean absolute error, and model training loss. |
| #pythonimplmentation | CS166: Modeling, simulation, and decision making. | In my project, I am reproducing the machine learning models proposed in the state-of-art paper. In my implementation, I am using some of the guidelines I learnt from this LO: using classes and functions, avoid using global variables, assign variable and function names that provide guidance to the reader, add comments and paragraphs in plain English to explain the type of the variables used and the purpose of each. |
| #interpretresults | CS166: Modeling, simulation, and decision making. | Using the results of the key metrics, we can draw conclusions on how each of the methods are performing relative to each other. |
| #optimalalogorithm | CS156: Machine Learning for Science and Profit | As we discussed in this LO, an optimal algorithm is the one that maintain “good” performance under the current constraints. We get to define what makes a performance good using the key metrics listed above. |
| #novelapplication | CS11b: Discrete Mathematical Systems | A novel application does not have to reinvent the wheel rather it aims to use the wheel creatively under some constraints. The second part of the paper discusses how we can create an alternative that takes the strength of all of the state-of-art algorithms to create another proposal. For example, one paper might have the strength of training time as the result of thorough pre-processing of the video. |
| #complexity | CS110: Computation: Solving Problems with Algorithms | Time and space complexity are key metrics that influence the model’s training/running time as a function of the size of the data. An efficient algorithm shall have a complexity of O(N), scale linearly. |
| #classification | CS156: Machine Learning for Science and Profit | The task we have on hand is classification. There are plenty of models that can do this task, i.e support vector machines, K-nearest neighbors. Given the complexity of the features of the data and the abundance of the later, using a neural network is a reasonable choice. |
| **#decisionbrief** |  |  |

# Appendix C: Potential sources of relevant support, advice, or feedback

Table 3. Points of contacts in the field.

|  |  |  |
| --- | --- | --- |
| Name | Relationship | Background |
| Gracie Ermi | Mentor | Machine Learning Software Engineering at Vulcan |
| Zach Barry | Mentor | Machine Learning/Data scientist at Novartis. |
| Sonia Kumar | Connector/mentor | Pharma D. |
| Phillip Wenig | Connector/ally | PhD. Machine learning Hasso Plattner Institute |

It is important to gain insights from experienced contacts that went through the same phase of ideation and project creation. For example, Gracie works as a computer-vision engineer and was my mentor for three months. Her insights were about the feasibility of the project and how I can plan ahead. Since we are connected on a personal level, she also gave me non-technical advice. For instance, she advised me to have a growth mindset and be able to handle my senior year emotional stress to make sure that I get my work done.

Professor Sterne is experienced with students who are pursuing a machine learning career after Minerva. Thus, his insights about the potential pros, cons, and constraints of the project helped me understand how to optimize project planning.

# Appendix D: Prototype

For this prototype, I implemented the approach used in (Ciftci, Demir & Yin, 2019).

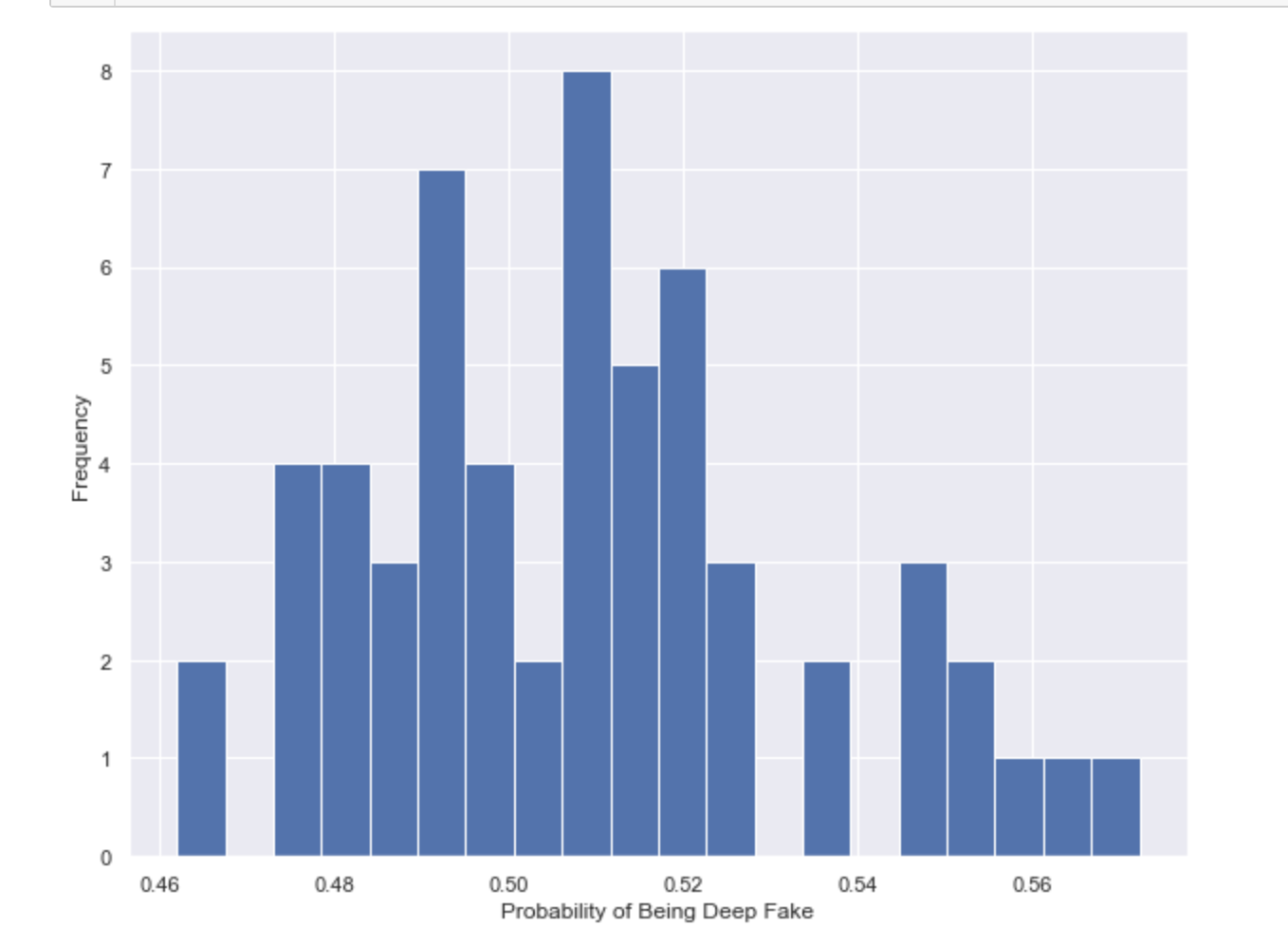


Figure 1. The distribution of the probability of a video being a deepfake. The test sample is 58. Full code: <https://github.com/asmaaalaa99/Minerva_Capstone/blob/master/Capstone%20prototype%20%20.ipynb>

# Appendix E: HCs and LOs applied in this proposal

1. #**qualitydeliverables**: I followed the template given and filled it with clear and well supported information
2. #**professionalism** (HC & LO): There is a clear and concise proposal that meets the minimum requirements and follows APA format
3. #**connect**: In Appendix C, I listed people I have a close relationship with and their expertise.
4. #**research**: For my background, I listed the papers that I am comparing, and stated the reason why I chose them.
5. #**metrics**: For creating, identifying, and applying appropriate metrics for the evaluation of your proposal, your project, and key objectives or deliverables. This can include HCs and LOs that can be used to evaluate your Capstone work products.
6. #**planningarchitecture**: I attached a screenshot of my board in Notion that shows the deadlines I created for myself. In addition to the full PDF version of the board.
7. #**evidencebased**: I used the sources effectively and stated their strengths and weaknesses. All of the sources are either peer-reviewed or published in reputable sources.
8. #**purpose**: I stated my rationale behind the deliverables: going to grad school after Minerva. Also, the level of the depth of the information that I provided matches the purpose of the assignment.