Pilot Operating Handbook Semantic Text Analysis

Project Description and Clarification

**The Boeing Company**

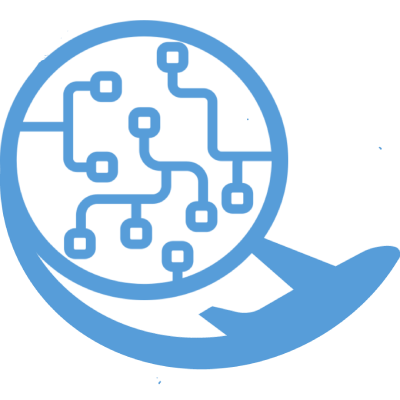


**The Autopilots**

Katie Peterson

Allison Lorphanpaibul

Weiren Lai

Daniel Harker

# Introduction

Artificial intelligence (AI) and machine learning (ML) have become the trends of the future; however, the cost of AI and ML is very high, especially in acquiring and maintaining real and production data. Hence, there is a need to be able to support training synthetic data engines like GANs (generative adversarial networks) to create synthetic data on-demand. These engines provide a fast and inexpensive way to produce data that is properly marked and tagged, unlike most real and production data.

Our goal is to develop a semantic text parsing algorithm that takes vocabulary from Boeing aircraft handbooks and stores each noun/noun phrase with its designated document name, year, product, location, and 1-50 sentences to demonstrate context for each noun. We expect the program to have a 60-80% quality after training the GAN [3].

# Background and Related Work

Our project will lie in the field of *Natural Language Processing*, how we teach machines to conceptualize human language. NLP has grown tremendously the last few years, along with AI and machine learning. The biggest issue with programming machines to understand our language is context as described in [1]. Any given word in the English language can have thousands of different meanings, depending on the context [1]. For example, the phrase ‘walk the plank’ will mean two different things given the context of a pilot handbook or a fictional book. Because of this problem, it is important to give a machine the information it needs so that it can eventually contextualize words itself.

The texts we will be generating will be written in Boeing’s *Simplified Technical English,* since that is their standard language for Pilot Handbooks. Therefore, it will be important that we provide our machine the necessary contextual information to understand this specific case.And while there are already many existing generative NLP algorithms, the given context and specificity of our algorithm will be something that has not been done before.

**Technical Skills**

For our project, we should read about already existing NLP models and how these models parse language into its different parts. Many models for specific NLP already exist. One example SCIBERT took the already existing BERT model to be specifically trained on scientific texts as explained in [2]. We can use the foundation that other models such as BERT have laid for our project.

Since we will be using GANs for our project, we should learn how to effectively use the GANs model and decide how we may need to tweak the model for our project’s needs.

We should also try to gain a better understanding of Boeing Simplified English. Understanding the context of the language used in the handbooks will help in training our classifier to parse the handbooks more effectively.

In terms of languages we may need to learn, our clients have suggested using either Python or Scala for this project. Everyone in our team already has experience with using Python’s scikit-learn library for data science, but we may have to learn how to use different libraries depending on the needs of the project. We will most likely need to try a variety of libraries, methods, and frameworks before finding the one that is best suited for our clients.

# Project Overview

**Project Problem**

The artificial intelligence and machine learning markets are rapidly growing, and are expected to continue to grow well into the future. However, testing solutions that utilize artificial intelligence or machine learning can be costly. These solutions often require vast amounts of data for testing, but acquiring and maintaining real and production data is expensive. Even once obtained, “data is often not marked or tagged sufficiently for [artificial intelligence] practitioners to understand the data” [3]. For this reason, the use of synthetic data in place of real production data has become a common practice.

As mentioned in the introduction, synthetic data engines, such as GANs, are used to create synthetic data on-demand. But before synthetic data engines can begin to produce data, they must be properly trained.

For a GAN that will produce text data, training often involves supplying it with linguistic rules and a set of vocabulary words. An important aspect of this is ensuring the GAN understands the meaning and context of the nouns in its vocabulary. In English, many nouns have different meanings depending on their context. It is thus important to ensure that when a GAN produces text data, it is using nouns appropriately.

We are aware that there are a few advantages and disadvantages to using GANs. GANs are hard to train and unstable. The generator and the discriminator (D/G) need to be well synchronized, but in actual training it is easy for discriminator to converge and generator to diverge. D/G training requires careful design. Additionally, mode collapse might become a problem. This means there may be missing patterns in the learning process of GANs, and the generator begins to degenerate, always generating the same sample points and not able to continue learning.

However, GANs can model the data distribution with a better view, where the image is sharper and clearer. In theory, GANs can train any kind of generator network. Other frameworks require the generator network to have some specific functional forms, an example being where the output layer is Gaussian. Also, there is no need to use Markov chain for repeated sampling, no need to infer in the learning process, no complicated variational lower bound, and it can avoid the difficult problem of approximate calculation of difficult probability [4].

**Objectives**

Our primary objective for this project is to aid the Boeing Company in its efforts to train GANs by creating a semantic text parsing algorithm that ultimately produces a constrained vocabulary based on the documents parsed.

The algorithm and accompanying components will be written in a language chosen by our team. It will then be applied to 3-10 publicly available aircraft documents to parse the nouns and noun phrases in the documents. Each document is a 100-300 page PDF and is written using Boeing Simplified English. After parsing, the nouns and noun phrases will be recorded in a MySQL database or Open Office spreadsheet, along with their corresponding document names, products, years, locations, and 1-50 sentences demonstrating their context.

**Goals & Outcomes**

In turn, the Boeing Company “will use the constrained vocabulary as a corpus to train GANs to produce synthetic paragraphs, sentences, and words.” As a result, the company expects “to find 60-80% quality after training the GAN, and almost instantaneous creation times. Real and production data with updated rules will be added to the vocabulary, delivering 90%+ quality test data” [3].

An additional goal and desired outcome of the project is for each team member to develop an understanding of the strategies, processes, and practices of work related to artificial intelligence and machine learning. The Boeing Company has expressed that these skills are critical for projects of the future, and thus those entering the workplace should possess these skills. As a result of completing this project, each team member will have applied these skills to real-world situations. This experience will be valuable for team members eventually entering the job market.

# Client and Stakeholder Identification and Preferences

**Primary Client**

**Boeing**

With our project’s problem statement coming from Boeing itself, it is clear that they will be our primary client. Through the use of our project’s final state, Boeing will be able to synthetically create sentences and paragraphs - saving them both time and money.

**Stakeholders**

**Commercial Airlines**

Hundreds of commercial airline companies, such as Delta and Alaska Airlines, depend on Boeing’s aircrafts. These companies are responsible for more than 10,000 of Boeing’s airplanes being flown.

**U.S. Military**

The U.S. military relies heavily on Boeing’s ability to produce high-quality aircrafts. Accounting for all military branches, Boeing currently has over 20 different types of aircrafts in service, totaling at more than 2,300 airplanes, helicopters, and drones.

**Pilots**

After Boeing, pilots will be the ones who are most affected by this project. With the need for pilots to man Boeing airplanes for the hundreds of commercial airlines, as well as man over 2,300 aircrafts for the U.S. military, there will be thousands of pilots directly affected.

**President of the United States**

In a more indirect manner, the President of the United States is a stakeholder of this project. Boeing is responsible for creating the VC-25, better known as Air Force One.

**Shipping Companies**

Many shipping companies use Boeing airplanes to send packages all around the world. A few notable companies include Amazon, FedEx, UPS, and USPS.

**U.S. Citizens**

Based on the fact that major shipping companies and commercial airlines are stakeholders, virtually every U.S. citizen will be a stakeholder as a result, granted they use these companies.

# Glossary

**Generative Adversarial Networks (GANs) -**  A popular machine learning model used to create synthetic data on-demand.

**Natural Language Processing (NLP) -** The field pertaining to how to program machines to understand human language.

**Bidirectional Encoder Representations from Transformers (BERT) -** A NLP model developed by Google with the intent to better understand contextual meanings in words.

**Simplified English -** The standard for written aerospace documents. Utilizes a restricted vocabulary and simple grammar rules.

# References

[1] D. Yse, “Your Guide to Natural Language Processing (NLP),” towardsdatascience.com*,* Jan. 15, 2019. [towardsdatascience.com/your-guide-to-natural-language-processing-nlp-48ea2511f6e1](https://towardsdatascience.com/your-guide-to-natural-language-processing-nlp-48ea2511f6e1). (accessed September 20, 2020).

[2] I. Beltagy, K. Lo. and Arman Cohan, “SCIBERT: A Pretrained Language Model for Scientific Text,” *arXiv,* Sep. 2019. <https://arxiv.org/pdf/1903.10676v3.pdf>.

[3] Rakshit Bhatt & Don Brancato, “Pilot Operating Handbook Semantic Text Analysis”, The Boeing Company, Washington, USA, 2020.

[4] 打不死的小强 . “一文看懂「生成对抗网络 - GAN」基本原理+10种典型算法+13种应用.” *产品经理的人工智能学习库*, 17 Dec. 2019, easyai.tech/ai-definition/gan/.