A black rectangle with a black background

Description automatically generated with low confidenceLogo, company name

Description automatically generated

DAEN 690

Project Report

Student Author Name 1

Student Author Name 2

Student Author Name 3

Student Author Name 4

Student Author Name 5

Student Author Name 6

Spring 2023

Report Title

**About the Cover**

Dr. Isaac Gang is an Associate Professor at the George Mason University College of Engineering and Computing, Volgenau School of Engineering, MS Data Analytics Engineering (DAEN) program.

He joined the DAEN faculty in the Fall of 2020 from Texas A&M University-Commerce (TAMUC) where he served as an Assistant Professor of Computer Science as well as the department’s Outreach Coordinator. Before coming to TAMUC, Dr. Gang was an Assistant Professor of Computer Science and Engineering at the University of Mary Hardin-Baylor (UMHB) and an Adjunct Professor of Computer Science at the University of Southern Mississippi’s School of Computing before joining UMHB.

Dr. Gang is a former DOE grant winner, former President and Board Member of the Association of Computer Educators in Texas (ACET), Industry Advisory Board (IAB) Coordinator, and the Director of CS For All.

His current and primary teaching responsibilities at Mason largely involves Data Analytics Engineering graduate courses along with a mix of CS and AIT graduate courses. He is an affiliate faculty member of GMU’s C4I & Cyber Center.

Dr. Gang’s primary research agenda involves Big Data Analytics (emphasis on data bias and data governance), Cyber Security (ransomware, steganography, and cyberbullying), and Image/Signal Processing.

Contents

Table of Contents

[Abstract 5](#_Toc125979012)

[Section 1: Problem Definition 7](#_Toc125979013)

[1.1 Background 7](#_Toc125979014)

[1.2 Problem Space 7](#_Toc125979015)

[1.3 Research 7](#_Toc125979016)

[1.4 Solution Space 8](#_Toc125979017)

[1.5 Project Objectives 8](#_Toc125979018)

[1.6 Primary User Stories 8](#_Toc125979019)

[1.7 Product Vision 9](#_Toc125979020)

[1.7.1 Scenario #1 9](#_Toc125979021)

[1.7.2 Scenario #2 9](#_Toc125979022)

[Section 2: Datasets 9](#_Toc125979023)

[2.1 Overview 9](#_Toc125979024)

[2.2 Field Descriptions 9](#_Toc125979025)

[2.3 Data Context 10](#_Toc125979026)

[2.4 Data Conditioning 11](#_Toc125979027)

[2.5 Data Quality Assessment 11](#_Toc125979028)

[2.6 Other Data Sources 11](#_Toc125979029)

[2.7 Storage Medium 11](#_Toc125979030)

[2.8 Storage Security 11](#_Toc125979031)

[2.9 Storage Costs 12](#_Toc125979032)

[Section 3: Algorithms & Analysis / ML Model Exploration & Selection 12](#_Toc125979033)

[3.1 Solution Approach 12](#_Toc125979034)

[3.1.1 Systems Architecture 12](#_Toc125979035)

[3.1.2 Systems Security 12](#_Toc125979036)

[3.1.3 Systems Data Flows 12](#_Toc125979037)

[3.1.4 Algorithms & Analysis 12](#_Toc125979038)

[3.2 Machine Learning 12](#_Toc125979039)

[3.2.1 Model Exploration 13](#_Toc125979040)

[3.2.2 Model Selection 13](#_Toc125979041)

[Section 4: Visualizations / ML Model Training, Evaluation, & Validation 13](#_Toc125979042)

[4.1 Overview 13](#_Toc125979043)

[4.2 Visualizations 13](#_Toc125979044)

[4.3 Machine Learning 13](#_Toc125979045)

[4.3.1 Model Training 13](#_Toc125979046)

[4.3.2 Model Evaluation 14](#_Toc125979047)

[4.3.3 Model Validation 14](#_Toc125979048)

[Section 5: Findings 14](#_Toc125979049)

[Section 6: Summary 14](#_Toc125979050)

[Section 7: Future Work 14](#_Toc125979051)

[Appendix A: Glossary 15](#_Toc125979052)

[Appendix B: GitHub Repository 16](#_Toc125979053)

[Overview 16](#_Toc125979054)

[GitHub Repository Link 16](#_Toc125979055)

[GitHub Repository Contents 16](#_Toc125979056)

[Appendix C: Risks 17](#_Toc125979057)

[Sprint 1 Risks 17](#_Toc125979058)

[Sprint 2 Risks 17](#_Toc125979059)

[Sprint 3 Risks 17](#_Toc125979060)

[Sprint 4 Risks 17](#_Toc125979061)

[Sprint 5 Risks 18](#_Toc125979062)

[Appendix D: Agile Development 19](#_Toc125979063)

[Scrum Methodology 19](#_Toc125979064)

[Sprint 1 Analysis 19](#_Toc125979065)

[Sprint 2 Analyis 19](#_Toc125979066)

[Sprint 3 Analysis 20](#_Toc125979067)

[Sprint 4 Analysis 20](#_Toc125979068)

[Sprint 5 Analysis 20](#_Toc125979069)

[Works Cited 21](#_Toc125979070)

**No table of figures entries found.**

**This page intentionally left blank**

Abstract

Abstract

**INSTRUCTIONS**

[NOTE: The project abstract is a separately graded assignment in the course. The final approved project abstract is to be copied word-for-word from the other assignment into this report.]

Write one paragraph of no more than 300 words that summarizes your project. Here are the typical kinds of information found in most abstracts which you should use as an outline as you develop your abstract.

1. The context or background information for your research; the general topic under study; the specific topic of your research.
2. The central questions or statement of the problem your research addresses.
3. What’s already known about this question, what previous research was conducted or shown.
4. The main reason(s), the exigency, the rationale, the goals for your research — why is it important to address these questions? Are you, for example, examining a new topic? Why is that topic worth examining? Are you filling a gap in previous research? Applying new methods to take a fresh look at existing ideas or data? Resolving a dispute within the literature in your field?
5. Your research and/or analytical methods.
6. Your main findings, results, or arguments.
7. The significance or implications of your findings or arguments.

Your abstract should be intelligible on its own, without a reader’s having to read your entire paper.

**DELETE THIS TEXT BOX AFTER YOU HAVE READ AND UNDERSTOOD THE INSTRUCTIONS.**

**This page intentionally left blank**

Report

# Problem Definition

## Background

**INSTRUCTIONS**

This section provides 4-5 pages of content for the reader with enough background information about the problem context that allows the paper to be standalone. In other words, assume the reader does not have a background for your project problem and you provide enough content in this section so the reader at a minimum can understand the problem space that you are going to discuss later. Terminology is important and guiding. For example, attempt to minimize the domain vocabulary and when introducing new terms make sure at a minimum, they are defined in Appendix A: Glossary.

Utilize figures, pictures, and tables since visualization are processed faster by the mind.

All references throughout the report follow the IEEE Citation Style.

**DELETE THIS TEXT BOX AFTER YOU HAVE READ AND UNDERSTOOD THE INSTRUCTIONS.**

## Problem Space

**INSTRUCTIONS**

This section describes the specific problem that you will attempt to solve completely or part of the problem. Note well that most project scope their solutions to a part of the problem space. Poor project teams usually fail to understand the problem and our eager to start implementing a solution to what they think is the problem. Most problems a very large and understanding the size of the problem mitigates the risk of foolhardy attempts that usually fail. This section should be about 2 pages and should show that the team understands the breadth and dept of the problem space.

Even with a well-defined sponsored project that may be scoped it is critical to provide the reader with the whole problem space in a concise a terse description, while providing the reader with a map of what part of the problem you are going to solve. This should allow the team to write the project extensions in the section 7 (Future Work). This is actually a lie since the team will discover many new path and areas of the problem space while working on the project.

**DELETE THIS TEXT BOX AFTER YOU HAVE READ AND UNDERSTOOD THE INSTRUCTIONS.**

## Research

**INSTRUCTIONS**

Document the research the team performed while either developing the solution or attempting to understand a solution path provided by the sponsor. This should be a summary of readings, Internet searches, collaborations, etc. Length of this section can vary but usually around 2-3 pages

**DELETE THIS TEXT BOX AFTER YOU HAVE READ AND UNDERSTOOD THE INSTRUCTIONS.**

## Solution Space

**INSTRUCTIONS**

Describe your solution approach. High level and provides your reader with an idea of approach.

Describe the solution space for the reader. For example: “Our system delivers value to its users when it accurately reports veracity scores for submitted articles. Users derive value from these scores when they feel more confident in their chosen news sources or avoid being misled when presented with fake news. We expect our system can help steer users to more authoritative news outlets by altering browsing behaviors.”

**DELETE THIS TEXT BOX AFTER YOU HAVE READ AND UNDERSTOOD THE INSTRUCTIONS.**

## Project Objectives

**INSTRUCTIONS**

Answer the following questions regarding the project objectives.

1. What does the team assume it will learn after finishing this project?
2. What does the team assume they will achieve as a solution when they finish this project?
3. What does the team assume it will achieve in terms of understanding about the problem after they finish this project?
4. What does the team assume it will provide in value as a product of this project work to the world, targeted group, etc.?

The above questions about the project objectives can be used to develop the primary user stories in section 1.6.

**DELETE THIS TEXT BOX AFTER YOU HAVE READ AND UNDERSTOOD THE INSTRUCTIONS.**

## Primary User Stories

**INSTRUCTIONS**

This story or stories explicitly state what the project is attempting to address: For example: Based on the user context and value proposition, we developed the following primary user story to guide our project:

“As a User, I want to submit an article to the Veracity System and receive a veracity score to know how much to trust a particular news article.”

**DELETE THIS TEXT BOX AFTER YOU HAVE READ AND UNDERSTOOD THE INSTRUCTIONS.**

## Product Vision

**INSTRUCTIONS**

Describe scenarios for why someone would want to use this solution.

* For:
* Who:
* The:
* Is a:
* That:
* Unlike:
* Our product:
* Caveats:

Provide at least two scenarios below for the project.

**DELETE THIS TEXT BOX AFTER YOU HAVE READ AND UNDERSTOOD THE INSTRUCTIONS.**

### Scenario #1

### Scenario #2

# Datasets

## Overview

**INSTRUCTIONS**

Provide a descriptive overview of your datasets.

**DELETE THIS TEXT BOX AFTER YOU HAVE READ AND UNDERSTOOD THE INSTRUCTIONS.**

## Field Descriptions

**INSTRUCTIONS**

Described your dataset field. Make sure you study the example below and you will more than likely expand these fields:

1. URL (Type: string) – The web address or Universal Resource Locator for the webpage that contained the news article. This includes the protocol (http or https), host name, and subdomain. Some URLs also include parameters (text following ‘?’) or named anchors (text following a ‘#’). Each URL can only be present once in the database, even if the webpage is not static over time.
2. Title (Type: string) – The title of the news article as parsed by the Newspaper 3K module. This field may be null (~150 articles in our dataset do not have titles).
3. Authors (Type: string) –The authors of the news article as parsed by the Newspaper 3K module. This field may be null (~23,000 articles do not have authors) and articles with multiple authors have their names joined with a comma into a single string. This field may also pick up descriptions of the author, including their titles and background.
4. Publication Date (Type: datetime) – The article publication date and time as parsed by the Newspaper 3K module. The datetime is displayed in ISO 8601 format (YYYY-MM-DD Thh:mm:ss+offset). Publish dates without specified times are assumed to be published at midnight. Publication dates with time information, but without a time zone listing, are assumed to be in Eastern Standard Time. This field is not allowed to be null.
5. Text (Type: string) – The text of an article as parsed by Newspaper 3K. This field may be null (~8,000 articles do not have text) as some news stories are delivered as only video, audio, or a picture. The mean word count for text is 538.9 across all news sources.
6. Tags (Type: string) – Article tags as determined by Newspaper 3K. These appear to be important (rare or “topicy”) words taken from the article text, not meta tags contained in the article’s HTML. Multiple tags are concatenated with a comma into a single string.

**DELETE THIS TEXT BOX AFTER YOU HAVE READ AND UNDERSTOOD THE INSTRUCTIONS.**

## Data Context

**INSTRUCTIONS**

Provide a description of the data context.

Data context is the set of circumstances that surround a collection of data. Capturing and interpreting context is a basic step in data analysis. Use of out-of-context data is a common source of errors in scientific research, business decisions, and professional advice.

In business analytics (BA), gathering context from external sources can provide useful information about events that have significance for the organization. Context for an unexplained surge in sales, for example, could be provided by pulling in data from news and social media as well as less obvious sources, such as weather over that period. Explored in context, it may be able to identify external causes for the increase, and that information might be used to guide future business decisions.

**DELETE THIS TEXT BOX AFTER YOU HAVE READ AND UNDERSTOOD THE INSTRUCTIONS.**

## Data Conditioning

**INSTRUCTIONS**

Describe the data conditioning required for each data set.

**DELETE THIS TEXT BOX AFTER YOU HAVE READ AND UNDERSTOOD THE INSTRUCTIONS.**

## Data Quality Assessment

**INSTRUCTIONS**

At a minimum you must assess your data sets with the following attributes:

* Completeness
* Uniqueness
* Accuracy
* Atomicity
* Conformity
* Overall Quality

**DELETE THIS TEXT BOX AFTER YOU HAVE READ AND UNDERSTOOD THE INSTRUCTIONS.**

## Other Data Sources

**INSTRUCTIONS**

If you are considered other data sources, however, you decided not to use these sources provide some reason why they were not utilized.

**DELETE THIS TEXT BOX AFTER YOU HAVE READ AND UNDERSTOOD THE INSTRUCTIONS.**

## Storage Medium

**INSTRUCTIONS**

Discuss the storage medium selected for the project data set storage.

**DELETE THIS TEXT BOX AFTER YOU HAVE READ AND UNDERSTOOD THE INSTRUCTIONS.**

## Storage Security

**INSTRUCTIONS**

Discuss the storage security required for the project data set storage.

**DELETE THIS TEXT BOX AFTER YOU HAVE READ AND UNDERSTOOD THE INSTRUCTIONS.**

## Storage Costs

**INSTRUCTIONS**

Discuss storage costs associated with the storage medium used for the project data set storeage,

**DELETE THIS TEXT BOX AFTER YOU HAVE READ AND UNDERSTOOD THE INSTRUCTIONS.**

# Algorithms & Analysis / ML Model Exploration & Selection

## Solution Approach

**INSTRUCTIONS**

Provide a detailed discussion of the solution approach. Include discussions on any of the following:

1. Systems Architecture
2. Systems Security
3. Systems Data Flows
4. Algorithms & Analysis
5. Machine Learning (delete this subsection for non-machine learning projects.

**DELETE THIS TEXT BOX AFTER YOU HAVE READ AND UNDERSTOOD THE INSTRUCTIONS.**

### Systems Architecture

### Systems Security

### Systems Data Flows

### Algorithms & Analysis

## Machine Learning

**INSTRUCTIONS**

For Machine Learning projects discuss the model exploration and selection process. Delete this report subsection for non-machine learning projects.

**DELETE THIS TEXT BOX AFTER YOU HAVE READ AND UNDERSTOOD THE INSTRUCTIONS.**

### Model Exploration

### Model Selection

# Visualizations / ML Model Training, Evaluation, & Validation

## Overview

**INSTRUCTIONS**

Provide an overview of what was accomplished during Sprint 4. Focus visualizations for non-machine learning projects.

**DELETE THIS TEXT BOX AFTER YOU HAVE READ AND UNDERSTOOD THE INSTRUCTIONS.**

## Visualizations

## Machine Learning

**INSTRUCTIONS**

For Machine Learning projects, discuss your approach to the following with respect to the ML Model:

1. Training,
2. Evaluation, and
3. Validation of the ML Model.

Delete this report subsection for non-machine learning projects.

**DELETE THIS TEXT BOX AFTER YOU HAVE READ AND UNDERSTOOD THE INSTRUCTIONS.**

### Model Training

### Model Evaluation

### Model Validation

# Findings

**INSTRUCTIONS**

Discuss the major findings of the project.

**DELETE THIS TEXT BOX AFTER YOU HAVE READ AND UNDERSTOOD THE INSTRUCTIONS.**

# Summary

**INSTRUCTIONS**

Summarize the overall project and results for the reader. What did you discover, prove, disprove, etc.

**DELETE THIS TEXT BOX AFTER YOU HAVE READ AND UNDERSTOOD THE INSTRUCTIONS.**

# Future Work

**INSTRUCTIONS**

This is critical section of the report. Propose future follow-on work or next step(s) for the project.

**DELETE THIS TEXT BOX AFTER YOU HAVE READ AND UNDERSTOOD THE INSTRUCTIONS.**

Appendix

Appendix A: Glossary

|  |  |
| --- | --- |
| Term | Definition |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

**INSTRUCTIONS**

Place all terms which require definitions in the Appendix A: Glossary.

**DELETE THIS TEXT BOX AFTER YOU HAVE READ AND UNDERSTOOD THE INSTRUCTIONS.**

Appendix B: GitHub Repository

Overview

**INSTRUCTIONS**

Provide a GitHub Link and the README.MD content. Do not just provide a link to the GitHub repository but provide a narrative paragraph which introduces the project. This section should mirror the look and feel of a well-documented professional GitHub site.

**DELETE THIS TEXT BOX AFTER YOU HAVE READ AND UNDERSTOOD THE INSTRUCTIONS.**

GitHub Repository Link

GitHub Repository Contents

Appendix C: Risks

Sprint 1 Risks

**INSTRUCTIONS**

Include the risk table associated with the Sprint. Below the risk table provide a narrative description of how the risks and mitigation plans were identified, what the team got correct, what the team could have done differently, how accurate was the team in identifying the risks, did the team encounter any unanticipated risks, etc. Think of this writeup as a “lessons learned” that you would like to pass along to any project team thinking of doing a similar project.

**DELETE THIS TEXT BOX AFTER YOU HAVE READ AND UNDERSTOOD THE INSTRUCTIONS.**

Sprint 2 Risks

**INSTRUCTIONS**

Include the risk table associated with the Sprint. Below the risk table provide a narrative description of how the risks and mitigation plans were identified, what the team got correct, what the team could have done differently, how accurate was the team in identifying the risks, did the team encounter any unanticipated risks, etc. Think of this writeup as a “lessons learned” that you would like to pass along to any project team thinking of doing a similar project.

**DELETE THIS TEXT BOX AFTER YOU HAVE READ AND UNDERSTOOD THE INSTRUCTIONS.**

Sprint 3 Risks

**INSTRUCTIONS**

Include the risk table associated with the Sprint. Below the risk table provide a narrative description of how the risks and mitigation plans were identified, what the team got correct, what the team could have done differently, how accurate was the team in identifying the risks, did the team encounter any unanticipated risks, etc. Think of this writeup as a “lessons learned” that you would like to pass along to any project team thinking of doing a similar project.

**DELETE THIS TEXT BOX AFTER YOU HAVE READ AND UNDERSTOOD THE INSTRUCTIONS.**

Sprint 4 Risks

**INSTRUCTIONS**

Include the risk table associated with the Sprint. Below the risk table provide a narrative description of how the risks and mitigation plans were identified, what the team got correct, what the team could have done differently, how accurate was the team in identifying the risks, did the team encounter any unanticipated risks, etc. Think of this writeup as a “lessons learned” that you would like to pass along to any project team thinking of doing a similar project.

**DELETE THIS TEXT BOX AFTER YOU HAVE READ AND UNDERSTOOD THE INSTRUCTIONS.**

Sprint 5 Risks

**INSTRUCTIONS**

Include the risk table associated with the Sprint. Below the risk table provide a narrative description of how the risks and mitigation plans were identified, what the team got correct, what the team could have done differently, how accurate was the team in identifying the risks, did the team encounter any unanticipated risks, etc. Think of this writeup as a “lessons learned” that you would like to pass along to any project team thinking of doing a similar project.

**DELETE THIS TEXT BOX AFTER YOU HAVE READ AND UNDERSTOOD THE INSTRUCTIONS.**

Appendix D: Agile Development

Scrum Methodology

**INSTRUCTIONS**

Provide a narrative of the team efforts in using a scrum methodology for a data analytics engineering project. Describe how easy/difficult was it to adapt to the Scrum methodology. Did the team conduct a daily scrum? If not, how often did the team conduct a scrum. Describe how easy/difficult it was to use the YouTrack tool to manage the project. Don’t be limited to just these questions. Think of this writeup as a “lessons learned” that you would like to pass along to any project team thinking of doing a similar project.

**DELETE THIS TEXT BOX AFTER YOU HAVE READ AND UNDERSTOOD THE INSTRUCTIONS.**

Sprint 1 Analysis

**INSTRUCTIONS**

Provide a narrative of the team’s efforts during this Sprint. Be sure to include – but not be limited to – how the team identified the User Stories, how well the team performed with the various tasks, how easy/difficult it was for the team to manage their activities during the Sprint, what did the team do correct, what could/should the team have done differently, etc. Think of this writeup as a “lessons learned” that you would like to pass along to any project team thinking of doing a similar project.

**DELETE THIS TEXT BOX AFTER YOU HAVE READ AND UNDERSTOOD THE INSTRUCTIONS.**

Sprint 2 Analyis

**INSTRUCTIONS**

Provide a narrative of the team’s efforts during this Sprint. Be sure to include – but not be limited to – how the team identified the User Stories, how well the team performed with the various tasks, how easy/difficult it was for the team to manage their activities during the Sprint, what did the team do correct, what could/should the team have done differently, etc. Think of this writeup as a “lessons learned” that you would like to pass along to any project team thinking of doing a similar project.

**DELETE THIS TEXT BOX AFTER YOU HAVE READ AND UNDERSTOOD THE INSTRUCTIONS.**

Sprint 3 Analysis

**INSTRUCTIONS**

Provide a narrative of the team’s efforts during this Sprint. Be sure to include – but not be limited to – how the team identified the User Stories, how well the team performed with the various tasks, how easy/difficult it was for the team to manage their activities during the Sprint, what did the team do correct, what could/should the team have done differently, etc. Think of this writeup as a “lessons learned” that you would like to pass along to any project team thinking of doing a similar project.

**DELETE THIS TEXT BOX AFTER YOU HAVE READ AND UNDERSTOOD THE INSTRUCTIONS.**

Sprint 4 Analysis

**INSTRUCTIONS**

Provide a narrative of the team’s efforts during this Sprint. Be sure to include – but not be limited to – how the team identified the User Stories, how well the team performed with the various tasks, how easy/difficult it was for the team to manage their activities during the Sprint, what did the team do correct, what could/should the team have done differently, etc. Think of this writeup as a “lessons learned” that you would like to pass along to any project team thinking of doing a similar project.

**DELETE THIS TEXT BOX AFTER YOU HAVE READ AND UNDERSTOOD THE INSTRUCTIONS.**

Sprint 5 Analysis

**INSTRUCTIONS**

Provide a narrative of the team’s efforts during this Sprint. Be sure to include – but not be limited to – how the team identified the User Stories, how well the team performed with the various tasks, how easy/difficult it was for the team to manage their activities during the Sprint, what did the team do correct, what could/should the team have done differently, etc. Think of this writeup as a “lessons learned” that you would like to pass along to any project team thinking of doing a similar project.

**DELETE THIS TEXT BOX AFTER YOU HAVE READ AND UNDERSTOOD THE INSTRUCTIONS.**

Reference

Works Cited

**There are no sources in the current document.**

**INSTRUCTIONS**

The References section of this document makes use of the Microsoft Word References feature to insert research citations by recording them directly into the document. All citations are to follow the IEEE citation format. Use the Bibliography drop down to have Microsoft Word dynamically create your Works Cited section

here in IEEE citation format.

To learn more about the IEEE Citation guidelines click on the document links below.

1. [IEEE-Reference-Guide.pdf](https://ieeeauthorcenter.ieee.org/wp-content/uploads/IEEE-Reference-Guide.pdf)
2. [IEEE Citation Guidelines2.doc (ieee-dataport.org)](https://ieee-dataport.org/sites/default/files/analysis/27/IEEE%20Citation%20Guidelines.pdf)

**DELETE THIS TEXT BOX AFTER YOU HAVE READ AND UNDERSTOOD THE INSTRUCTIONS.**

**This page intentionally left blank**