Syracuse University School of Information Studies
M.S. Applied Data Science

Portfolio Milestone

David Caspers
SUID: 50619985
dcaspers@syr.edu / caspersdavid@gmail.com

David Caspers
SUID: 50619985

Table of Contents

David Caspers
SUID: 50619985

Introduction

This paper reflects on my experience in the Syracuse University Master of Science in Applied Data Science (ADS) program, linking my applied projects and research to the program's learning objectives. Enrolling in this program has been part of an effort to switch careers. Therefore, my personal learning objectives for this program has been to develop both a broad foundation in analytic and big data processing skills with specialized theoretical and applied expertise in natural language processing (NLP) and deep learning**.** Through my coursework and projects, I developed skills that align with the core learning outcomes of the ADS program, particularly in:

1. Data Collection & Storage: Leveraging Python, Spark, SQL, and cloud-based tools to collect, store, and manage large-scale structured and unstructured data.
2. Data Analysis & Model Development: Applying machine learning, NLP, and deep learning techniques to extract insights across business, societal, and political contexts.
3. Predictive Modeling & Visualization: Building data-driven models and visualizations to create actionable insights.
4. Programming & Data Science Tools: Proficiency in Python, R, and SQL with detailed knowledge of specialized modeling and data processing toolkits such as TensorFlow [1], Keras [2], Scikit-learn [3], Pandas, NumPy, and NLP libraries to develop and deploy AI models.
5. Communication & Decision-Making: Translating technical findings into actionable insights for both technical and non-technical stakeholders.
6. AI Ethics & Responsible Modeling: Ensuring fairness, transparency, and bias mitigation in AI applications.

To build a strong theoretical foundation, I selected secondary core courses aligned with two key specializations: AI & Deep Learning and Language Analytics. In AI and deep learning, I focused on advanced predictive models, particularly in computer vision and NLP applications. In language analytics, I explored text mining and NLP techniques to analyze political rhetoric, sentiment analysis, and automated content classification. These courses provided the theoretical grounding necessary to understand and implement these concepts in real-world applications. More specifically, Applied Machine Learning (IST 707) introduced fundamental machine learning and statistical modeling techniques for structured datasets, while Natural Language Processing (IST 664) covered core NLP methodologies for working with unstructured text. Deep Learning in Practice (IST 691) provided insight into deep learning architectures for computer vision and NLP, and Data Administration Concepts and Database Management (IST 659) established a strong understanding of database management, schema design, and SQL optimization for practical applications.

The projects in this portfolio showcase my ability to apply these skills to real-world problems. In my MySQL Invoice Management System project [4], I designed a database solution for a small landscaping business, improving invoice generation and management through optimized SQL queries. In Geospatial Vessel Activity Analysis [5], I developed a pipeline to detect illegal, unregulated, and unreported (IUU) fishing using vessel movement data, employing clustering techniques such as K-Means and DBSCAN, as well as supervised learning models like Random

Forest and SVMs. In Deep Learning for Lung Cancer Detection [6], I implemented a Convolutional Neural Network (CNN) and Vision Transformer [7] model to analyze histopathology images, leveraging attention mechanisms to improve model interpretability. Finally, in Opinion Mining for Political Ideology [8], I scraped and curated a dataset of 37,000+ political statements, fine-tuning BERT models to assess where they fall in the political spectrum relative to their colleagues.

The projects included in this portfolio were selected based on their alignment with the ADS program's learning outcomes and my career objectives in AI and NLP. Each project demonstrates:

- Technical capability: Use of AI, deep learning, and NLP techniques deployable to real-world production systems.
- Real-world impact: Addressing practical challenges in political analysis, healthcare, and geospatial intelligence.
- Continuous learning: Evolution of skills across multiple domains of data science.

Together, these projects reflect my ability to apply data science principles beyond the classroom, preparing me for industry roles in AI, machine learning, and NLP applications.

Database Model for Quote Management System

a. Project Description

My final project for IST 659, Data Administration Concepts and Database Management, implemented a Landscaping Quote Management System [4] to streamline job estimation, invoicing, and payment processing for landscaping businesses. This system was designed to help businesses generate accurate, professional quotes while tracking customer information, material costs, service rates, employee wages, and payments. By improving pricing accuracy and financial tracking, the system ensures profitability and operational efficiency.

In developing and implementing the database, the scope was structured to support multiple customers, quotes, invoices, and payments, with each quote including line items for materials, labor, and service costs. The system required organizing structured data to automate financial processes while ensuring data integrity and relational consistency. Conceptual and logical models were developed to organize relationships between customers, quotes, invoices, materials, payments, and discounts (Fig. 1).

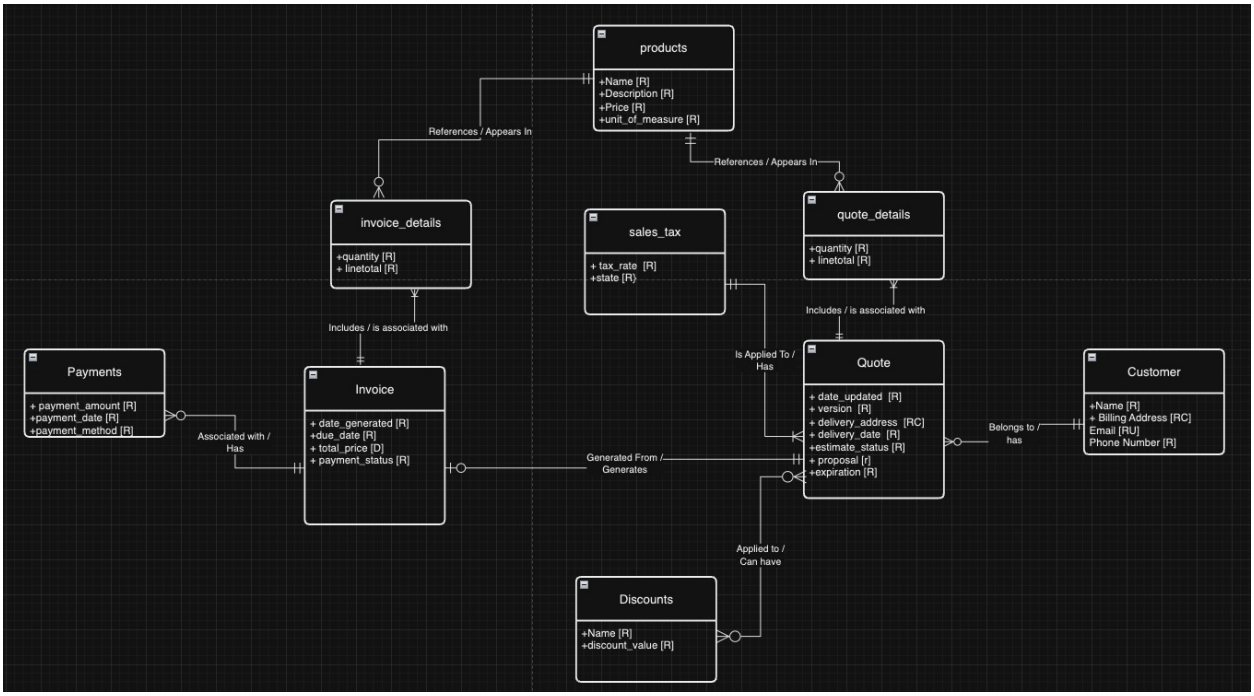

Figure 1 - Database Conceptual Model

Database implementation was performed using MySQL, with tables designed to store customer details, quotes, invoice records, payments, and service pricing. Data population and queries were executed using SQL scripts, ensuring that the system could automatically generate invoices from approved quotes, enforce financial accuracy by only accepting full payments, and apply dynamic pricing rules such as sales tax and discounts. Stored procedures were implemented to process payments, generate invoices from quotes, and update pricing dynamically based on business rules.

SQL queries were developed to support reports that provide actionable insights, such as tracking pending quotes, identifying overdue invoices, analyzing revenue trends, and monitoring the most frequently requested services (Fig. 2). While creating a user interface was out of scope for the project, a mock-up was created in PowerPoint to illustrate the potential reporting functionality that would require the implemented database infrastructure and queries created as part of this project.

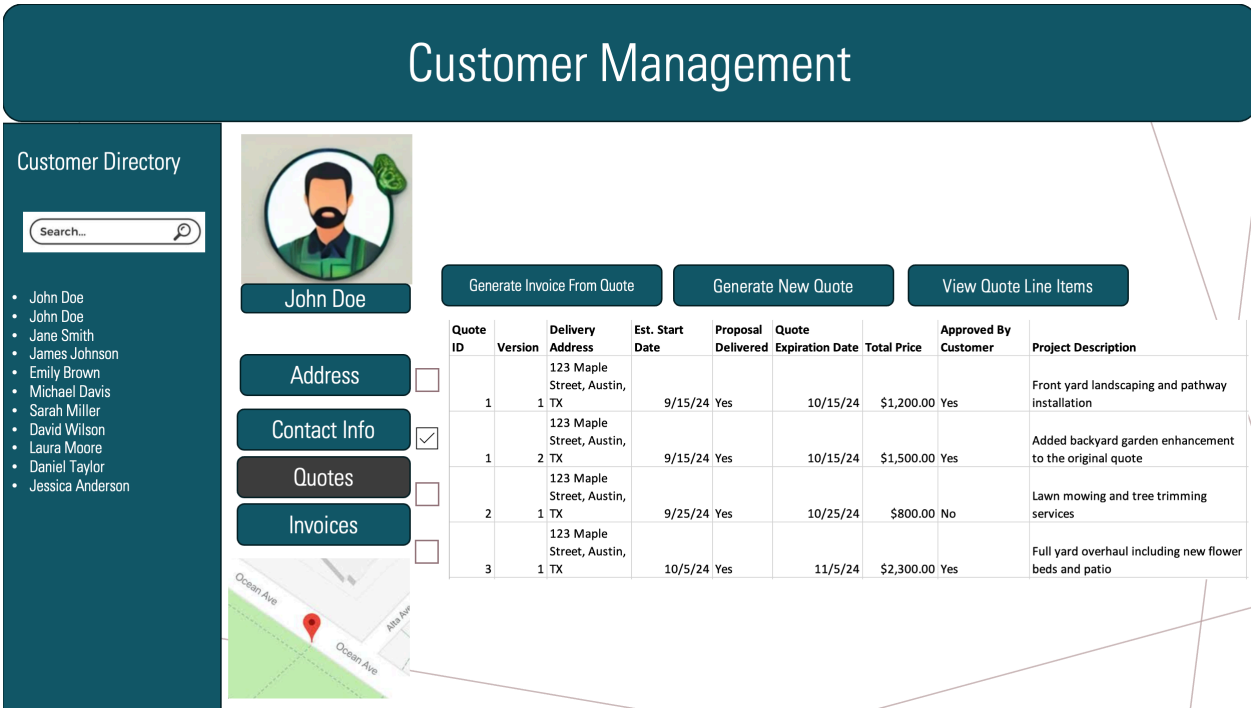

Figure 2 - Mock-up of Database Application

b. Project Reflection and Key Takeaways

The development of the Quote Management system reinforced key Applied Data Science competencies, particularly in Data Collection & Storage, Programming & Data Science Tools, and Communication & Decision-Making. Designing relational schemas, normalizing data, and implementing structured queries provided hands-on experience in organizing and accessing data efficiently – fundamental skills for any data-driven project.

A significant takeaway was the value of a formal design methodology. Systematically collecting and documenting business requirements, translating them into database models, and implementing a structured database ensured that every design decision was auditable and aligned with business needs. I was surprised by the complexity of implementing even a simple invoicing system—ensuring data integrity, optimizing query performance, and modeling relationships for accurate financial calculations was no trivial task. While this system was not built for machine learning, the structured design approach I learned provides a strong foundation for building scalable analytics and big-data applications.

Moving forward, the structured design principles from this project extend beyond database development into broader project design and implementation. This experience led me to adopt a more standardized approach, including using the Cookiecutter Data Science framework [9] for organizing my projects. As a result, I have refactored my lung cancer and political ideology projects in this portfolio detailed in the following sections to align with this structured format, ensuring consistency, scalability, and maintainability.

Additionally, this project sparked my interest in managing data systems in the cloud. To deepen my understanding, I pursued and earned the AWS Certified Solutions Architect – Associate certification [10] in August 2024 and the AWS Certified Machine Learning Specialty certification [11] in January 2025. These certifications have strengthened my ability to design scalable, secure cloud-based architectures, a crucial skill for deploying real-world data science applications. Moving forward, I plan to explore integrating cloud-based solutions for efficient data storage, processing, and model deployment in future projects.

Combating Illegal, Unreported, and Unregulated Fishing

a. Project Description

Illegal, Unreported, and Unregulated (IUU) fishing is a global issue that threatens marine ecosystems, depletes fish stocks, and causes billions in economic losses annually. This project aimed to detect fishing activity using Automatic Identification System (AIS) data from Global Fishing Watch [12], which tracks vessel movements through geospatial signals [5]. By leveraging machine learning models, the goal was to classify whether a vessel was actively engaged in fishing or simply in transit, providing a foundation for identifying potential illegal fishing operations. The dataset contained over 4.3 million AIS signals from 49 trawling vessels, with key features such as speed, geolocation, heading, and distance from shore/port. A data preprocessing pipeline was developed to clean sensor errors, remove outliers, and introduce lag variables, capturing temporal patterns that could improve classification performance.

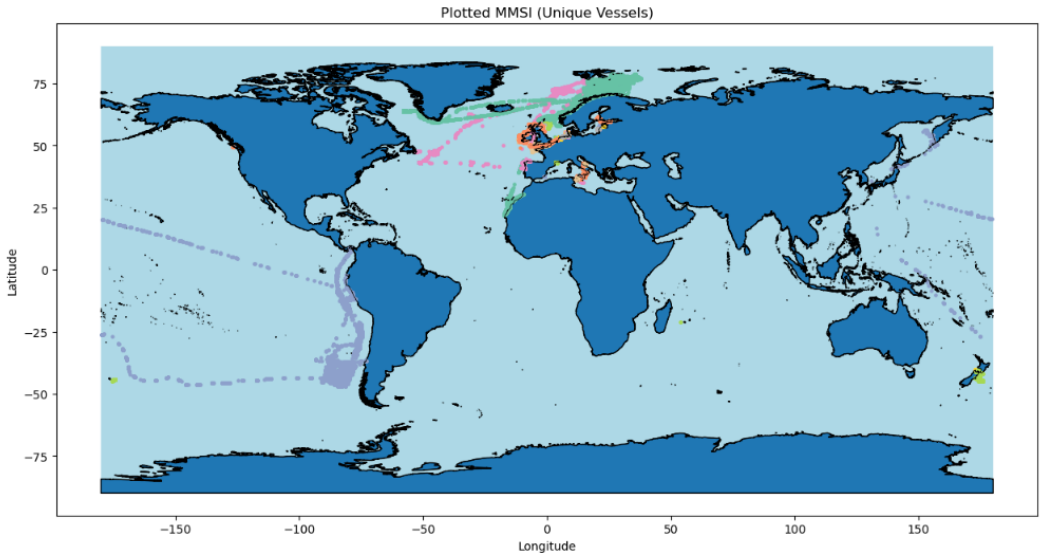

Figure 3 - Plot of AIS Data, Color Coded by Unique Ships

A series of machine learning models, including Random Forest, Support Vector Machines (SVM), and unsupervised clustering techniques (K-Means & DBSCAN), were used to explore the data and determine the most effective approach. One of the key challenges in this project was differentiating fishing from transit, as vessels often exhibit similar movement patterns when slowing down near fishing zones or ports. Data cleaning was necessary, as AIS data frequently contained sensor errors and outliers that could mislead the model. Additionally, accounting for

the temporal information in the dataset by ensuring that training and test data were properly stratified by time was necessary to avoid data leakage, which could artificially inflate model performance. To address these challenges, time series cross-validation was used to ensure that the model was evaluated on unseen vessel data, providing a more realistic assessment of its performance.

For the clustering methods DBSCAN proved the most effective at ignoring noise (erroneous signals) and managed to create clusters of similar fishing activity rather than only clustering based on geographical proximity, which included both transiting and actively fishing vessels. For classification, the Random Forest model with lag variables outperformed all others, achieving 88% accuracy in distinguishing fishing from non-fishing activities. The most influential features included vessel speed and distance from shore, with the inclusion of lagged variables significantly improving predictive power. Support Vector Machines also underperformed, likely due to the complexity of vessel behavior and the difficulty in finding a clear decision boundary.

This analysis showcases a method to create a scalable and automated solution for monitoring commercial fishing activity. With further refinement, this approach could support government agencies, conservation organizations, and law enforcement in identifying and investigating IUU fishing operations. Useful future work to build on this work could include integrating real-time AIS feeds, improving detection accuracy in overlapping transit/fishing areas, and incorporating additional maritime data sources to enhance classification reliability. This project highlights the potential of machine learning in maritime intelligence, offering a data-driven tool to combat illegal and unsustainable fishing practices.

b. Project Reflection and Key Takeaways

This project reinforced key data science competencies, particularly in data analysis and model development, predictive modeling, programming in Python, and communication and decision-making. One of the key takeaways was the challenges associated with working with real-world datasets, especially the complexities of data preparation and cleaning. The Automatic Identification System (AIS) data required extensive preprocessing to remove sensor errors, handle missing values, and address outliers that could have negatively impacted model performance.

Another significant challenge was working with time-series data, which required the application of lag variables and time-series specific validation techniques. Implementing time-series cross-validation ensured that the model was tested on unseen future data rather than introducing artificial accuracy through random splits. The ability to apply these structured methodologies deepened my understanding of how to handle temporal dependencies in machine learning models and reinforced the importance of selecting the appropriate evaluation strategies when working with sequential data.

The project also required analyzing geographic data, which presented unique challenges related to coordinate systems, time zones, and spatial relationships. While geospatial analysis aided in understanding vessel movement patterns, handling different coordinate reference systems and ensuring consistency in geographic data representation required careful attention. Certain aspects

of visualization were simplified by mapping latitude and longitude points, but the complexity of interpreting and aligning geospatial features across different time zones and datasets introduced additional challenges.

Ultimately, this exercise further prepared me for a career in data science by reinforcing key program learning outcomes in data analysis and model development, predictive modeling, and programming. The experience of cleaning complex datasets and applying machine learning techniques to real-world problems strengthened my ability to extract meaningful insights from structured and unstructured data. Additionally, working with time-series and geographic data deepened my understanding of the specialized tools and techniques required for handling non-traditional datasets. Finally, by systematically documenting my approach and aligning my model results with real-world objectives, this project provided valuable experience in articulating insights clearly and making complex data more actionable. It reinforced my ability to present findings in a way that informs strategic decisions and drives meaningful outcomes.

Detecting Lung Cancer from Histopathological Images

a. Project Description

Lung cancer remains one of the deadliest cancers worldwide, making early and accurate detection critical for improving patient outcomes [13]. This project developed a deep learning model to classify histopathological images of lung tissue into three categories: adenocarcinoma, squamous cell carcinoma, and normal tissue (Figure 4) [6]. The goal was to compare the performance of a Convolutional Neural Network (CNN) and a Vision Transformer [7] (ViT) in automating this classification process.

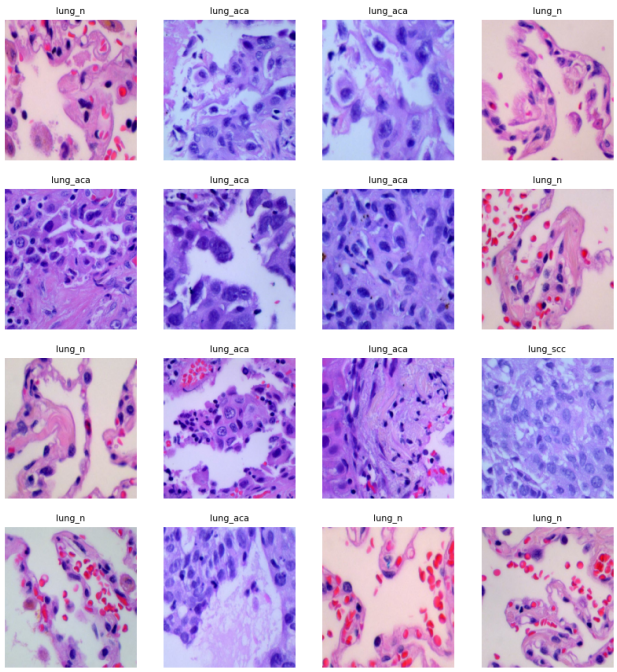

Figure 4 - Sample of Training Images

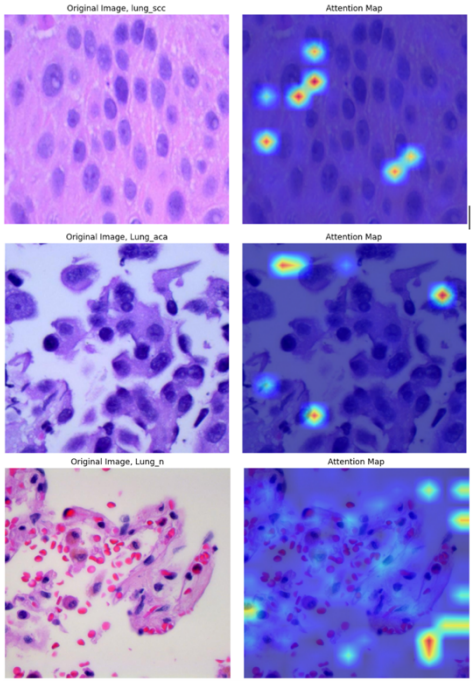

Figure 5 - Visualized Attention Maps

CNNs are well-suited for extracting local features in medical images, while ViTs incorporate attention mechanisms to capture spatial relationships and long-range dependencies across an image. By leveraging these strengths, the ViT model improved classification accuracy and reduced false negatives, which is critical in a medical setting. The final model was fine-tuned using transfer learning and integrated with attention heatmaps (Figure 5) to provide visual explanations of its decisions, enhancing interpretability for medical professionals.

b. Project Reflection and Key Takeaways

One of the most significant challenges was handling 15,000 high-resolution histopathological images, which required memory-efficient processing techniques such as lazy evaluation and batch processing to prevent system overload. In retrospect after taking additional courses in big data processing, incorporating distributed computing frameworks like Apache Spark could have further optimized data handling.

This project also reinforced my understanding of Vision Transformers and their advantages in medical image analysis. Unlike CNNs, ViTs rely on self-attention mechanisms to retain global context, making them well-suited for distinguishing subtle histopathological patterns. Utilizing TensorFlow / Keras and pretrained models from Hugging Face, I gained practical experience in fine-tuning and deploying advanced transformer architectures for medical classification tasks.

A key takeaway was the importance of explainability in AI-driven diagnostics. Deep learning models, particularly transformers, are often considered black boxes, making clinical adoption challenging. The attention heatmaps addressed this by highlighting the most critical regions

influencing classification decisions. This approach enhances trust and interpretability, ensuring AI-assisted diagnostics align with the decision-making processes of healthcare professionals.

This project gave me practical experience in deep learning model development, data preprocessing, and structured evaluation of AI models, aligning with the ADS program's learning outcomes, including Data Collection & Storage, Predictive Modeling & Visualization, and AI Ethics & Responsible Modeling. It also introduced me to new techniques for improving model interpretability, which is particularly useful for projects in industries where explainability is essential.

Inferring Where Politicians Rank on Political Spectrum From Public Statements

a. Project Description

This project explores how natural language processing (NLP) techniques can be used to infer where a politician falls on the political spectrum relative to their peers based on their public statements. By analyzing opinion-based rhetoric, the goal was to determine how well a politician's language reflects their ideological stance. The project involved two key tasks:

1. Extracting opinion statements using a fine-tuned DistilBERT [14] model to classify subjective (opinion-based) and objective (fact-based) sentences.
2. Ranking politicians ideologically by converting their subjective statements into vector embeddings and projecting them onto an ideological axis defined by known ideological baselines from the Limited Government Scorecard [16].

The corpus of political statements was scraped from the web using the VoteSmart API [15], collecting public statements, speeches, and interviews from 430 U.S. politicians, with over 37,000 subjective sentences identified for analysis. The methodology relied on fine-tuning DistilBERT for sentence classification, ensuring that only opinion-based statements were retained for ideological ranking. These sentences were then embedded into high-dimensional space, and each politician's collective rhetoric was represented by averaging their sentence embeddings. The ranking approach compared these embeddings against reference ideological baselines, producing a ranking that was evaluated against expert-curated ideological scores. Figure 6 illustrates the approach used to create the rankings but only shows the embeddings in two dimensional space to allow it be plotted. The model's performance metrics—including a Spearman correlation of 0.607 and Kendall Tau of 0.403—indicate a moderate but meaningful relationship between predicted rankings and expert evaluations.

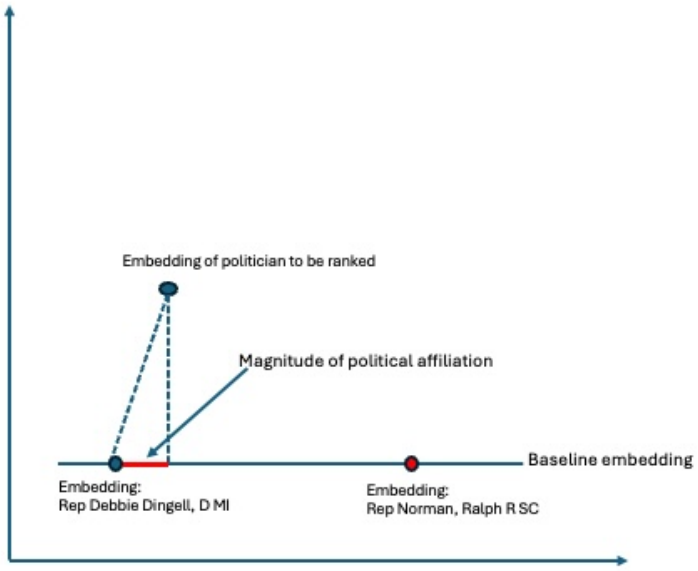

Figure 4 - Illustration of Ranking Methodology

While the model effectively captured ideological trends, certain limitations were observed. Politicians with limited or less ideologically distinct speech data were more difficult to classify accurately. Future improvements could include enhancing feature representation by incorporating additional linguistic and contextual signals, refining ranking methods to better capture ideological nuances, and experimenting with alternative distance metrics or ensemble approaches. By applying machine learning to large-scale political text analysis, this work demonstrates the potential for automating political insight generation.

b. Project Reflection and Key Takeaways

This project strengthened key data science learning outcomes, particularly in data collection and storage, data analysis and model development, predictive modeling, programming in Python, and communication and decision-making. Scraping political statements from the VoteSmart API and handling a large corpus of unstructured text data reinforced the importance of efficient data acquisition, preprocessing, and transformation techniques. Managing 37,000+ subjective statements required structuring text data for downstream NLP tasks, improving my ability to work with real-world, large-scale text data pipelines.

From a technical perspective, this project enhanced my NLP modeling skills by fine-tuning DistilBERT for sentence classification and using sentence embeddings for ranking politicians ideologically. Evaluating multiple classification approaches, including Naïve Bayes and TF-IDF-based models, reinforced the importance of model selection and feature engineering. The embedding-based ranking methodology, which projected politicians onto an ideological axis, demonstrated how predictive modeling techniques can extend beyond standard classification tasks. These experiences directly support the applied data science learning objectives, ensuring I

can develop and implement NLP models, structure large-scale text data, and apply ranking methodologies to real-world problems.

Reflection & Future Growth

Throughout my coursework in the Master of Science in Applied Data Science (ADS) program at Syracuse University, I have tackled diverse projects that collectively demonstrate my ability to apply data science methodologies to complex, real-world problems. These projects span multiple domains, including business applications, geospatial intelligence, healthcare, and political analysis, showcasing my adaptability and technical expertise across structured and unstructured data.

Each project reinforced fundamental data science concepts:

- Data Collection & Storage: I worked with SQL databases, API-driven data extraction, and unstructured text processing, ensuring efficient data management for downstream analysis.
- Predictive Modeling & Machine Learning: From decision trees and support vector machines to deep learning models like CNNs and Vision Transformers, I gained experience selecting, tuning, and interpreting machine learning models for various tasks.
- Natural Language Processing (NLP): Fine-tuning BERT-based models for opinion mining in political rhetoric enabled me to explore the complexities of language modeling and sentiment analysis.
- Geospatial Data Analysis: In detecting illegal fishing activity, I applied clustering algorithms and supervised classification methods to analyze time-series location data.
- Deep Learning & Computer Vision: Implementing deep learning architectures to classify histopathological images enhanced my ability to work with complex imaging data and attention mechanisms.
- Data-Driven Decision-Making & Communication: Each project required me to translate technical insights into actionable findings, preparing me to effectively communicate with both technical and non-technical stakeholders.

The breadth and depth of these projects illustrate my capacity to navigate the entire data science pipeline—from data acquisition and cleaning to model development, evaluation, and deployment. These experiences have prepared me to approach professional challenges with structured methodologies and innovative problem-solving techniques.

While these projects have provided me with a strong foundation in data science, there remain areas where I seek to deepen my expertise:

- Scalability & Big Data Processing: Although I have worked with large datasets, further exploration of distributed computing frameworks like Apache Spark and cloud-based machine learning workflows will be beneficial.

- MLOps & Model Deployment: Gaining hands-on experience with model monitoring, continuous integration and continuous development (CI/CD) pipelines, and production deployment of machine learning models is crucial for real-world applications.
- Explainability & Ethical AI: While I have incorporated attention mechanisms and model interpretation techniques, continuous learning in AI fairness, bias mitigation, and responsible AI deployment will be essential as the field evolves.
- Advanced NLP Techniques: Expanding my knowledge of transformer architectures, including Retrieval Augmented Generation (RAG) and Large Language Model Agents, will enable me to push the boundaries of NLP applications.

To maintain my growth in the field, I plan to engage in continuous learning through:

- Participation in data science competitions (e.g., Kaggle challenges) to refine my practical skills.
- Gaining additional professional certifications focused on cloud computing and big-data processing such as Databricks, Apache Kafka, and AWS Professional Solutions Architect.
- Attending industry conferences and workshops on AI advancements and emerging technologies.
- Pursuing research opportunities and collaborative projects that bridge my expertise in AI, deep learning, and political text analysis.

My academic journey in the ADS program has equipped me with the technical and analytical skills necessary to excel in data science roles across industries. The ability to structure data pipelines, develop machine learning models, and communicate insights will be valuable in professional settings, including AI-driven business analytics, healthcare informatics, and policy research.

These projects have not only honed my technical skills but also reinforced the importance of interdisciplinary collaboration, ethical considerations, and stakeholder engagement. The applied nature of my work has prepared me for roles that require problem-solving, innovation, and adaptability in leveraging data-driven solutions for business and societal impact.

Moving forward, I am confident that my expertise in machine learning, deep learning, and NLP—combined with my continuous pursuit of knowledge—will position me to contribute meaningfully to data science advancements in both industry and research settings.

References

[1] M. Abadi et al., "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. Available: https://www.tensorflow.org/

[2] F. Chollet, "Keras," *GitHub Repository*, 2015. Available: https://github.com/keras-team/keras.

[3] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011. Available: https://scikit-learn.org/stable/

[4] D. Caspers, "Database Model for Quote Management System," Syracuse University, 2024. Available: https://github.com/caspersd/misc_projects/tree/main/Database_Quote_Management_Application

[5] D. Caspers, "Combating Illegal, Unreported, and Unregulated Fishing," Syracuse University, 2024. Available: https://github.com/caspersd/Illegal_Fishing_Detection/tree/main

[6] D. Caspers, "Detecting Lung Cancer from Histopathological Images," Syracuse University, 2024. Available: *https://github.com/caspersd/detecting_lung_cancer*

[7] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, Oct. 2020. Available: https://arxiv.org/abs/2010.11929

[8] D. Caspers, "Inferring Ideological Alignment of Politicians from Their Public Statements," Syracuse University, 2024. Available: *https://github.com/caspersd/political_ideology_detection*

[9] D. T. Abernathy et al., "Cookiecutter Data Science," DrivenData, 2017. Available: https://cookiecutter-data-science.drivendata.org/.

[10] Amazon Web Services, "AWS Certified Solutions Architect – Associate," AWS Training & Certification, 2024. Available: https://aws.amazon.com/certification/certified-solutions-architect-associate/.

[11] Amazon Web Services, "AWS Certified Machine Learning – Specialty," AWS Training & Certification, 2024. Available: https://aws.amazon.com/certification/certified-machine-learning-specialty/.

[12] Global Fishing Watch, "Public Training Data v1," Global Fishing Watch. [Online]. Available: https://globalfishingwatch.org/data-download/datasets/public-training-data-v1. [Accessed: Mar. 2025].

[13] World Health Organization, "Cancer," Dec. 2024. Available: https://www.who.int/news-room/fact-sheets/detail/cancer

[14] V. Sanh et al., "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter," arXiv preprint arXiv:1910.01108, Oct. 2019. Available: https://arxiv.org/abs/1910.01108

[15] VoteSmart, "VoteSmart API," 2023. Available: https://www.votesmart.org/votesmart-api

[16] Institute for Legislative Analysis, "Limited Government Scorecard," 2022. Available: https://scorecard.limitedgov.org/guides/US-Guide-2022.pdf