# Classifying ETDs

CS4624: Multimedia, Hypertext, and Information Access
Professor Edward Fox
Virginia Tech, Blacksburg VA 24061
17 May, 2023

By -
Vedant Shah
Reema Daniel
Vaishali Ramesh
Mihir Gathani

# Outline

# Team Member Roles

- **Mihir Gathani**
  - Team Lead, R&D (Data Preprocessing work), Website Building, Annotation.
- **Vedant Shah**
  - R&D (Text Classification models), Point of Contact with Client, Dataset Development.
- **Reema Daniel**
  - Point of Contact with Dr. Fox, Annotation, and Website Development and Integration.
- **Vaishali Ramesh**
  - Noting minutes meeting, Website Development, and Annotation.

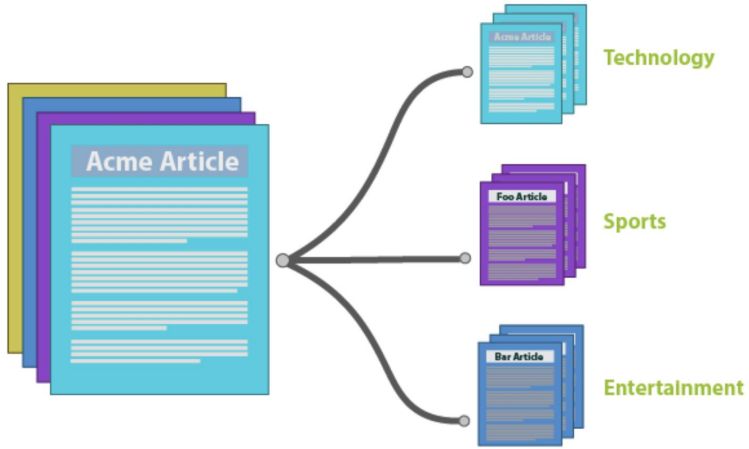
# Client

- Name: Bipasha Banerjee
- Occupation: PhD student on GRA (Graduate Research Assistant) advised by Dr. Fox
- Area of Expertise: Machine Learning, Natural Language Processing
- Research Focus: Information Retrieval from book length documents.

Note: Image approved by the client

# What's the Project about?

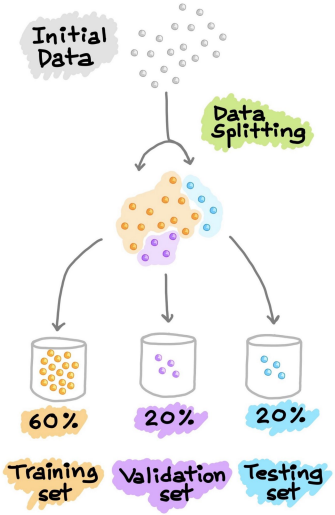
- Classification - Process of identifying and grouping objects or ideas into predetermined categories.
- Electronic Theses and Dissertations (ETDs) - are documents explicating and expressing research in a form suitable for both machine archives and worldwide retrieval.
- Purpose - Automate the process of classification of ETDs.



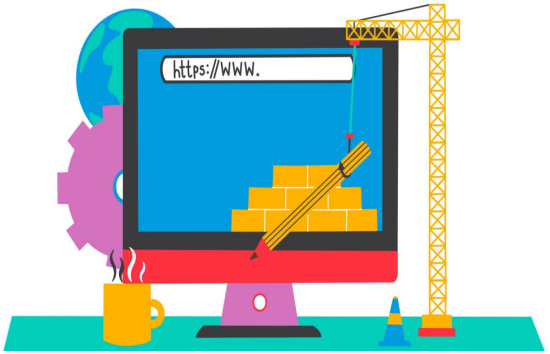
Text Classification

# Deliverables


Gold standard dataset

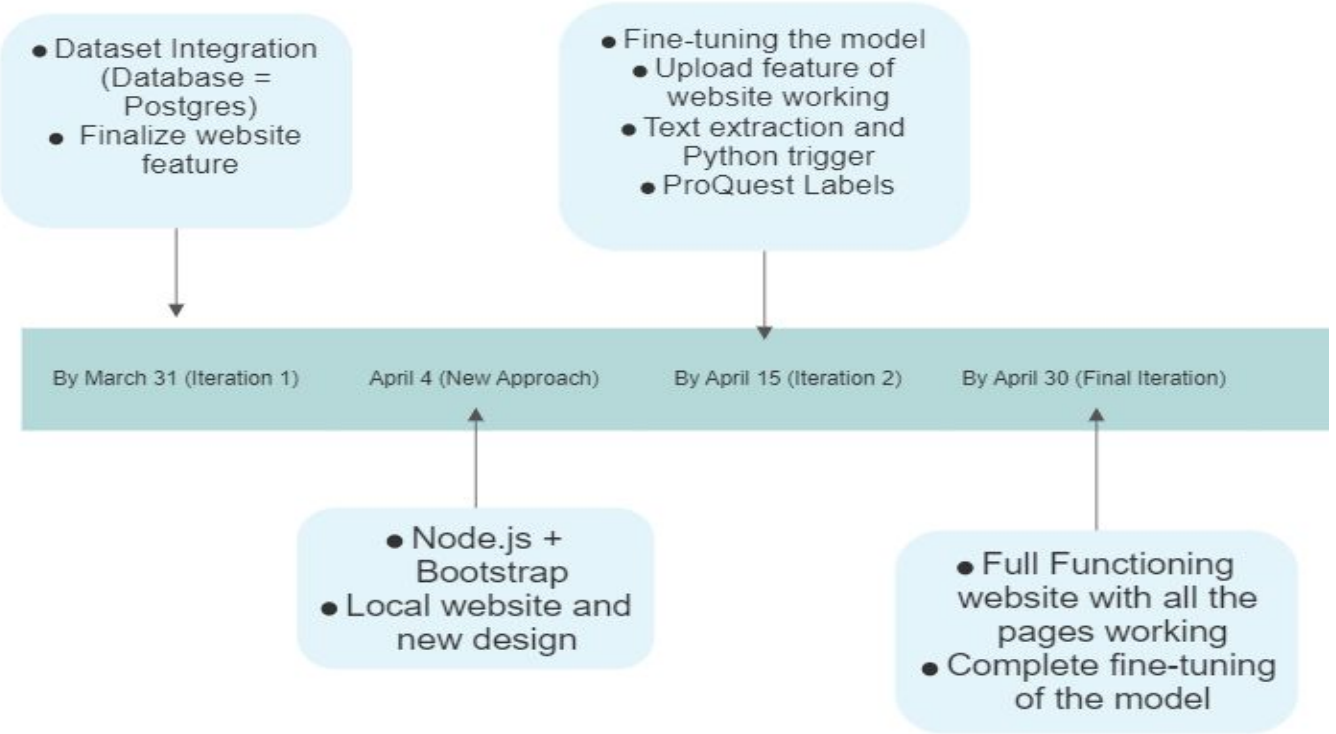



High Precision Text Classification Model




Interactive website for visualization

# Timeline


• Dataset Integration (Database = Postgres)
• Finalize website feature
• Fine-tuning the model
• Upload feature of website working
• Text extraction and Python trigger
• ProQuest Labels
By March 31 (Iteration 1)
April 4 (New Approach)
By April 15 (Iteration 2)
By April 30 (Final Iteration)
• Node.js + Bootstrap
• Local website and new design
• Full Functioning website with all the pages working
• Complete fine-tuning of the model

# Task Completed : Dataset

| | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | id | title | author | advisor | year | abstract | university | degree | URI | department | discipline | language | schooltype | oadsclassifier | |
| 2 | 1436 | Plasma Diagnostics and Plasma-Sur | Titus, Monica Joy | Graves, David B; | 2010 | The semic | ucb | | https://escholarship.org/uc/item/0hn5z4f1 | | | eng | REGULAR | 0 | |
| 3 | 1437 | Declarative Systems | Condie, Tyson | Hellerstein, Josep | 2011 | Building sy | ucb | | https://escholarship.org/uc/item/0sn1r9st | | | eng | REGULAR | 0 | |
| 4 | 1438 | Portrait of the Rugged Individualist | Horberg, Elizabeth Jane | Keltner, Dacher; | 2010 | Emotions | ucb | | https://escholarship.org/uc/item/0v37d9g2 | | | eng | REGULAR | 0 | |
| 5 | 1439 | Essays in Empirical Macroeconomi | Nelson Mondragon, Johr | Gorodnichenko, Y | 2015 | This disser | ucb | | https://escholarship.org/uc/item/0wh0h5bj | | | eng | REGULAR | 0 | |
| 6 | 1440 | Control and Trajectory Generation | Swift, Timothy Alan | Kazerooni, Homa | 2011 | There are | ucb | | https://escholarship.org/uc/item/0xc9q3b6 | | | eng | REGULAR | 0 | |
| 7 | 1441 | Errors as a Productive Context for | Leveille Buchanan, Nicol | Saxe, Geoffrey; | 2016 | How do te | ucb | | https://escholarship.org/uc/item/0zz775v7 | | | eng | REGULAR | 0 | |

Original Metadata

| | | Unnamed: 0 | id | year | year | abstract | university | degree | URI | department | discipline | schooltype | Updated_Dept | Stem_NonStem |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 0 | 29687 | 238148 | mining metals, mining minds: | 2021.0 | this dissertation examines the | vanderbilt university graduate school | PhD | http://hdl.handle.net/1803 | history | history | REGULAR | history | Non Stem |
| 3 | 1 | 30023 | 238577 | hard time in the new deal: raci | 2004.0 | "hard time in the new deal" bri and native american inmates, t | the university of texas at austin | History | http://hdl.handle.net/2152 | history | history | REGULAR | history | Non Stem |
| 4 | 2 | 30034 | 238588 | william jenkins, business elites | 2008.0 | this is a biographical case stud | the university of texas at austin | History | http://hdl.handle.net/2152 | history | history | REGULAR | history | Non Stem |
| 5 | 3 | 30154 | 238733 | creating a national passion: fo | 2004.0 | the introduction of football to s | the university of texas at austin | History | http://hdl.handle.net/2152 | history | history | REGULAR | history | Non Stem |
| 6 | 4 | 30156 | 238735 | negotiating a slave regime: fre | 2004.0 | in my dissertation, i investigate | the university of texas at austin | History | http://hdl.handle.net/2152 | history | history | REGULAR | history | Non Stem |
| 7 | 5 | 30158 | 238738 | "the bukom boys" : subculture | 2003.0 | in the early 1960s, kpanlogo, a | the university of texas at austin | History | http://hdl.handle.net/2152 | history | history | REGULAR | history | Non Stem |
| 8 | 6 | 30296 | 238936 | dueling perceptions : british ar | 2007.0 | this study is an attempt to asc | the university of texas at austin | History | http://hdl.handle.net/2152 | history | history | REGULAR | history | Non Stem |
| 9 | 7 | 30308 | 238948 | conquest and resistance in co | 2007.0 | in a 1963 article, "epic and cou that two different languages, c ignorance of each other: musli | the university of texas at austin | History | http://hdl.handle.net/2152 | history | history | REGULAR | history | Non Stem |

Final Metadata

| | Dept | Code | Label | Category |
|---|---|---|---|---|
| **0** | african american studies | 296 | AREA, ETHNIC, AND GENDER STUDIES | Arts, Business, Education, Humanities, and Soc... |
| **1** | african studies | 293 | AREA, ETHNIC, AND GENDER STUDIES | Arts, Business, Education, Humanities, and Soc... |
| **2** | american studies | 323 | AREA, ETHNIC, AND GENDER STUDIES | Arts, Business, Education, Humanities, and Soc... |
| **3** | asian american studies | 343 | AREA, ETHNIC, AND GENDER STUDIES | Arts, Business, Education, Humanities, and Soc... |
| **4** | asian studies | 342 | AREA, ETHNIC, AND GENDER STUDIES | Arts, Business, Education, Humanities, and Soc... |
| **...** | ... | ... | ... | ... |
| **424** | statistics | 463 | MATHEMATICAL AND PHYSICAL SCIENCES | Behavioral, Natural, and Physical Sciences |
| **425** | statistical physics | 219 | MATHEMATICAL AND PHYSICAL SCIENCES | Behavioral, Natural, and Physical Sciences |
| **426** | thermodynamics | 348 | MATHEMATICAL AND PHYSICAL SCIENCES | Behavioral, Natural, and Physical Sciences |
| **427** | theoretical mathematics | 642 | MATHEMATICAL AND PHYSICAL SCIENCES | Behavioral, Natural, and Physical Sciences |
| **428** | theoretical physics | 753 | MATHEMATICAL AND PHYSICAL SCIENCES | Behavioral, Natural, and Physical Sciences |

429 rows × 4 columns

ProQuest Dataset

| ETD_Depts | ProQuest_Depts | similarity |
|---|---|---|
| history | history | 1.0 |
| communication studies | communication | 0.707107 |
| kinesiology and health education | health education | 0.707107 |
| theatre and dance | dance | 0.707107 |
| physics | physics | 1.0 |
| ... | ... | ... |
| public health | public health | 1.0 |
| sport management | management | 1.0 |
| computer science and engineering | computer engineering | 0.707107 |
| molecular genetics | genetics | 0.707107 |
| graduate school | school counseling | 0.707107 |

Cosine Matches

| ETD_Depts | ProQuest_Depts | similarity |
|---|---|---|
| civil architectural and environmental engineering | architectural engineering | 0.632456 |
| biological science | information science | 0.5 |
| chemistry and biochemistry | biochemistry | 0.57735 |
| government | african american studies | 0.0 |
| advertising | african american studies | 0.0 |
| ... | ... | ... |
| criminal justice and criminology | criminology | 0.57735 |
| brand and media strategy | area planning and development | 0.5 |
| <-- please select one --> | african american studies | 0.0 |
| speech-language pathology | language | 0.57735 |
| evolution, ecology and organismal biology | biology | 0.5 |

Cosine Misses

| | ETD_Depts | ProQuest_Depts | Code | Label | Category |
|---|---|---|---|---|---|
| 0 | electrical and computer engineering | electrical engineering,computer engineering | 544,464 | ENGINEERING | Behavioral, Natural, and Physical Sciences |
| 1 | psychology | psychology | 621 | BEHAVIORAL SCIENCES | Behavioral, Natural, and Physical Sciences |
| 2 | mechanical engineering | mechanical engineering | 548 | ENGINEERING | Behavioral, Natural, and Physical Sciences |
| 3 | computer science | computer science | 984 | ENGINEERING | Behavioral, Natural, and Physical Sciences |
| 4 | civil architectural and environmental engineering | civil engineering,architectural engineering,en... | 543,462,775 | ENGINEERING \| ARCHITECTURE \| ENVIRONMENTAL SCI... | Behavioral, Natural, and Physical Sciences \| A... |
| 5 | chemistry | chemistry | 485 | MATHEMATICAL AND PHYSICAL SCIENCES | Behavioral, Natural, and Physical Sciences |
| 6 | physics | physics | 605 | MATHEMATICAL AND PHYSICAL SCIENCES | Behavioral, Natural, and Physical Sciences |

Final Mappings Matches

# Task Completed : Text Classification Models

- Developed Code for Fine-Tuning LLMs.
- Developed Inference Code for Generating Probabilistic Outputs.

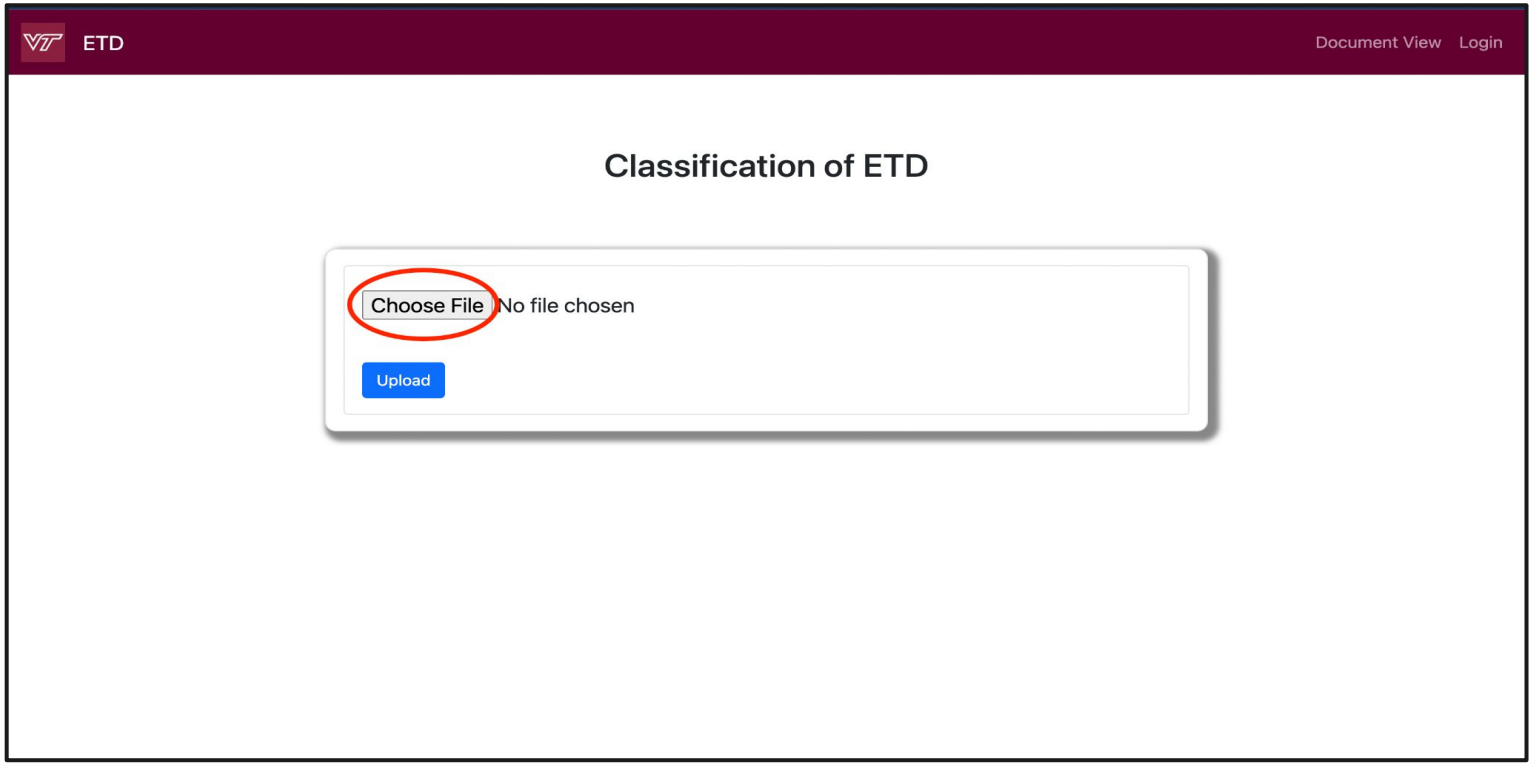
```
!python /content/run.py "gpt2" "/content/sample_data/chapter3_parsed.txt"
```

```
Text File Path is Valid, Proceeding ...
Using GPU: Tesla T4
[['computer science', 'electrical and computer engineering', 'journalism']]
[[0.9925368428230286, 0.002030385425314307, 0.0007162086549215019]]
```

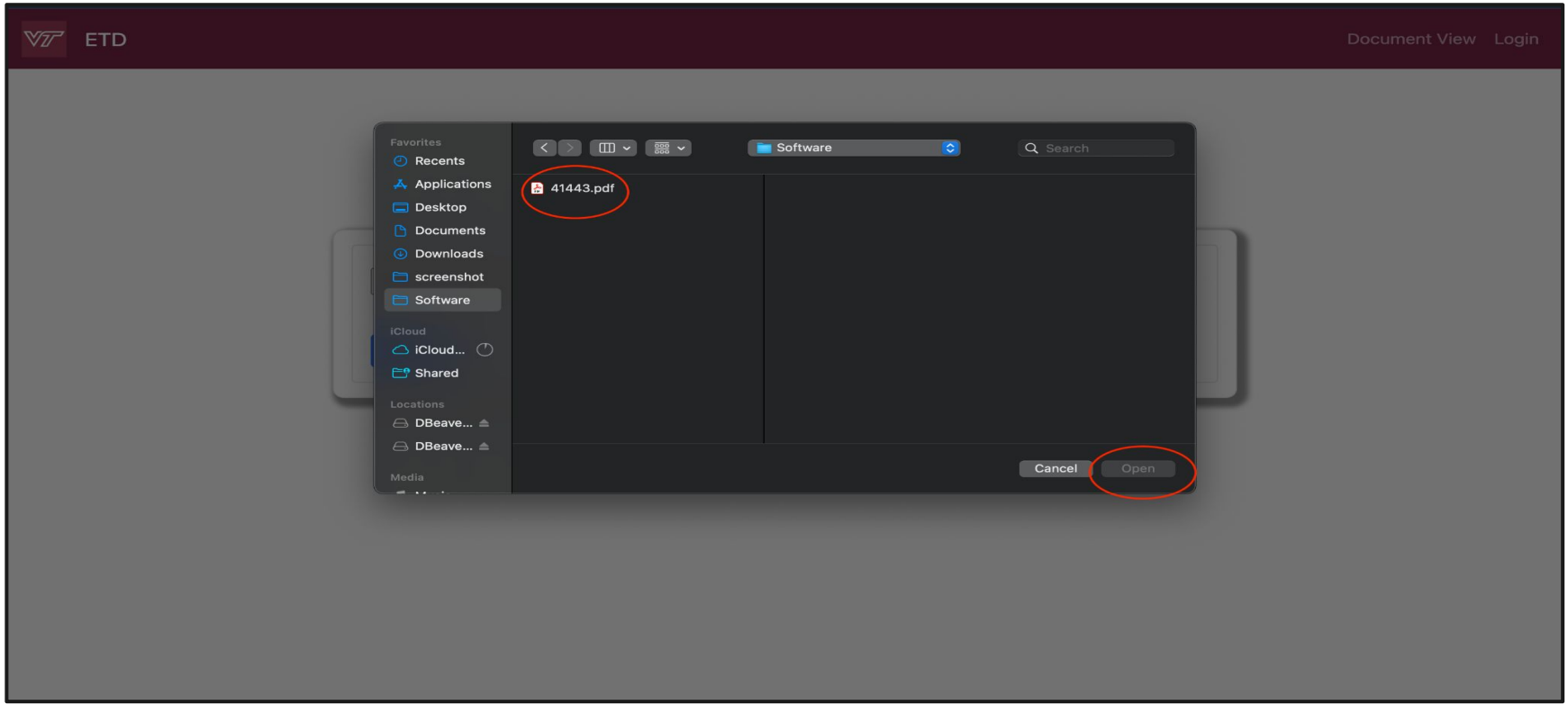
| MODEL NAME | METRICS<br>Train: Title + Abstract | Train: Full ETD Text |
|---|---|---|
| SciBERT | F1-Score: 72.4% | 78.4% |
| BioBERT | F1-Score: 72.4% | 74% |
| GPT-2 | F1-Score: 73.8% | - |
| BigBird-Pegasus | F1-Score: 73.4% | - |

Table of Baseline Predictions for different Fine-Tuned Models.

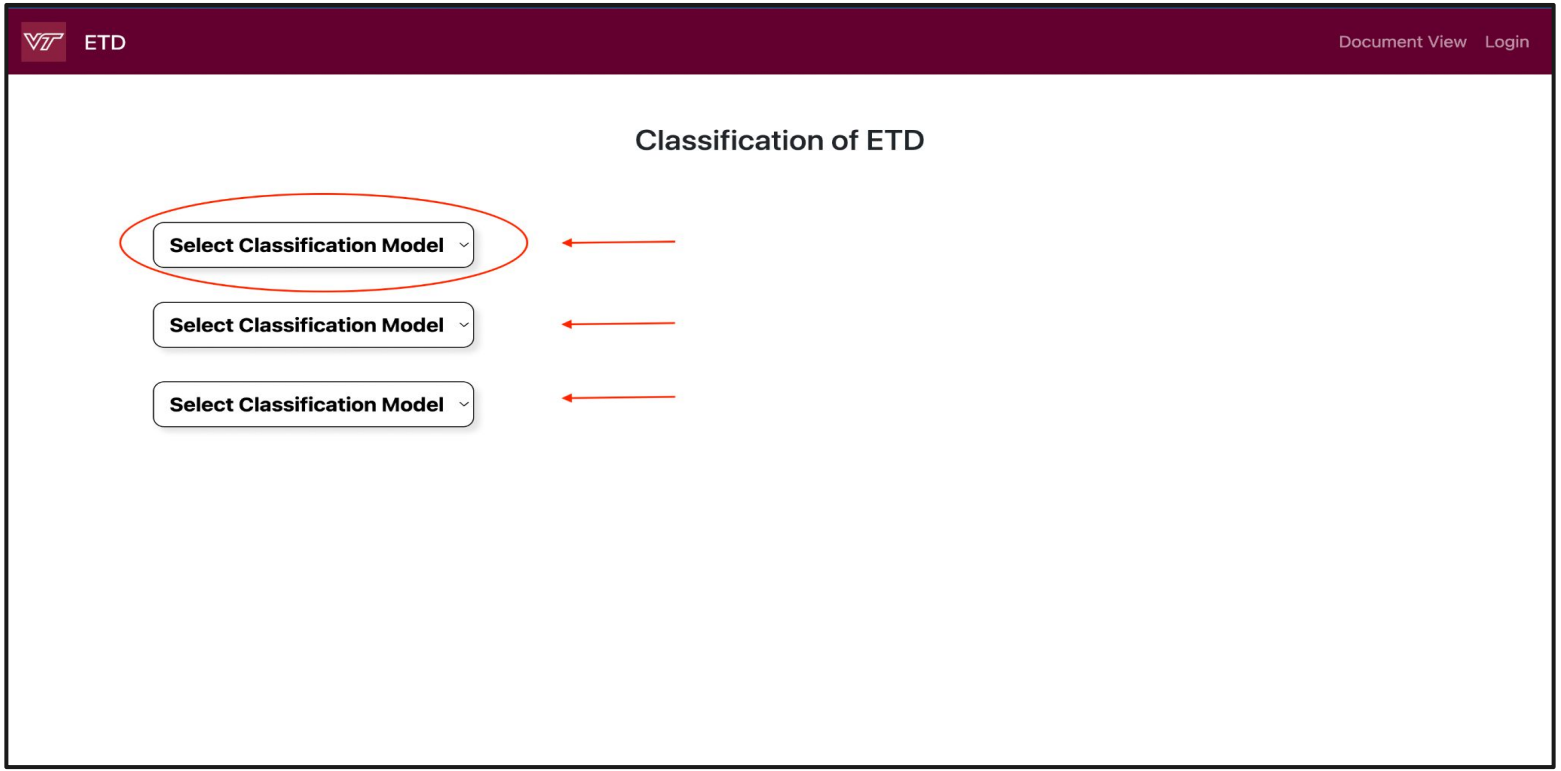
# Task Completed - Website Development



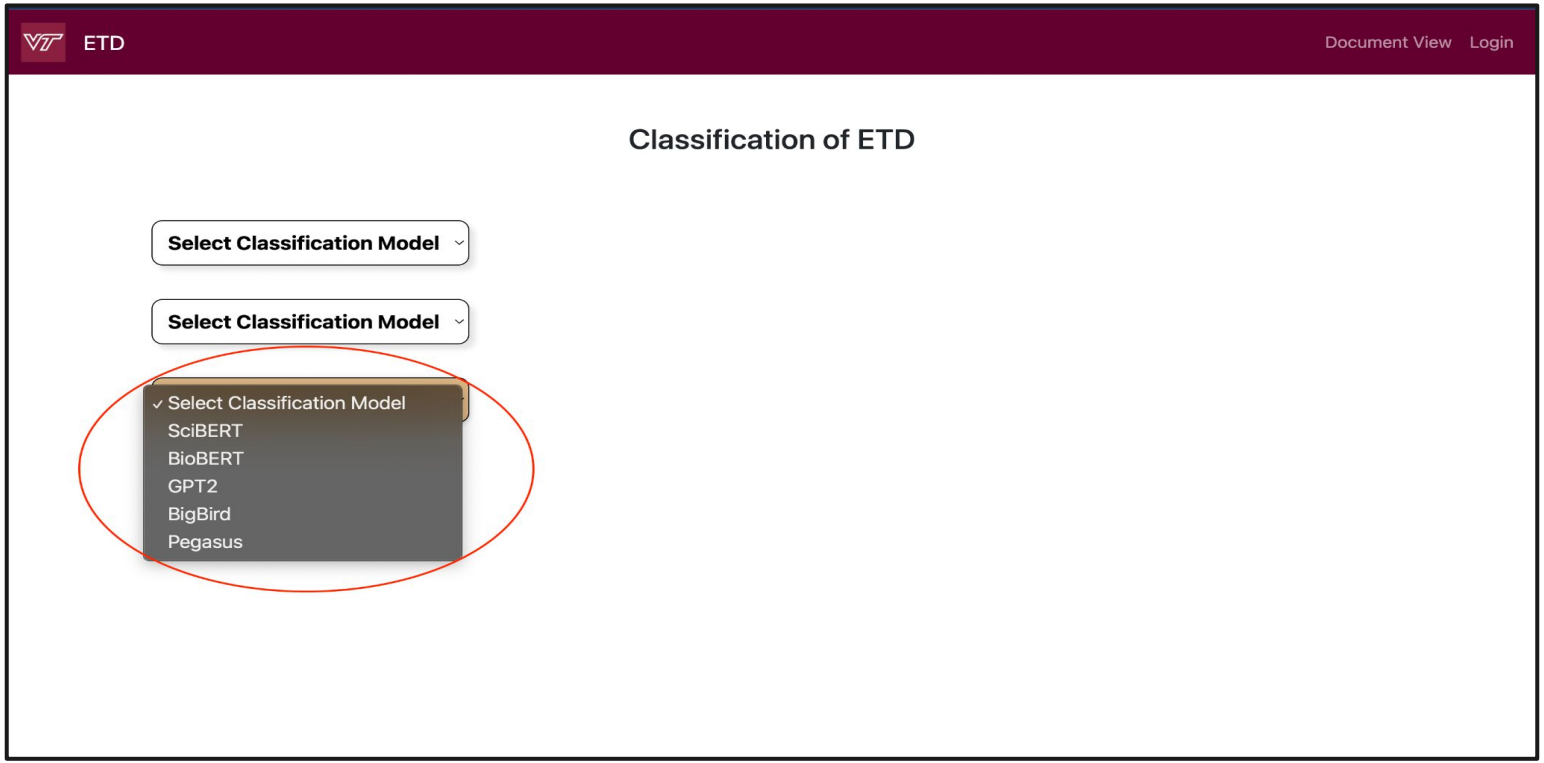
# Upload Feature



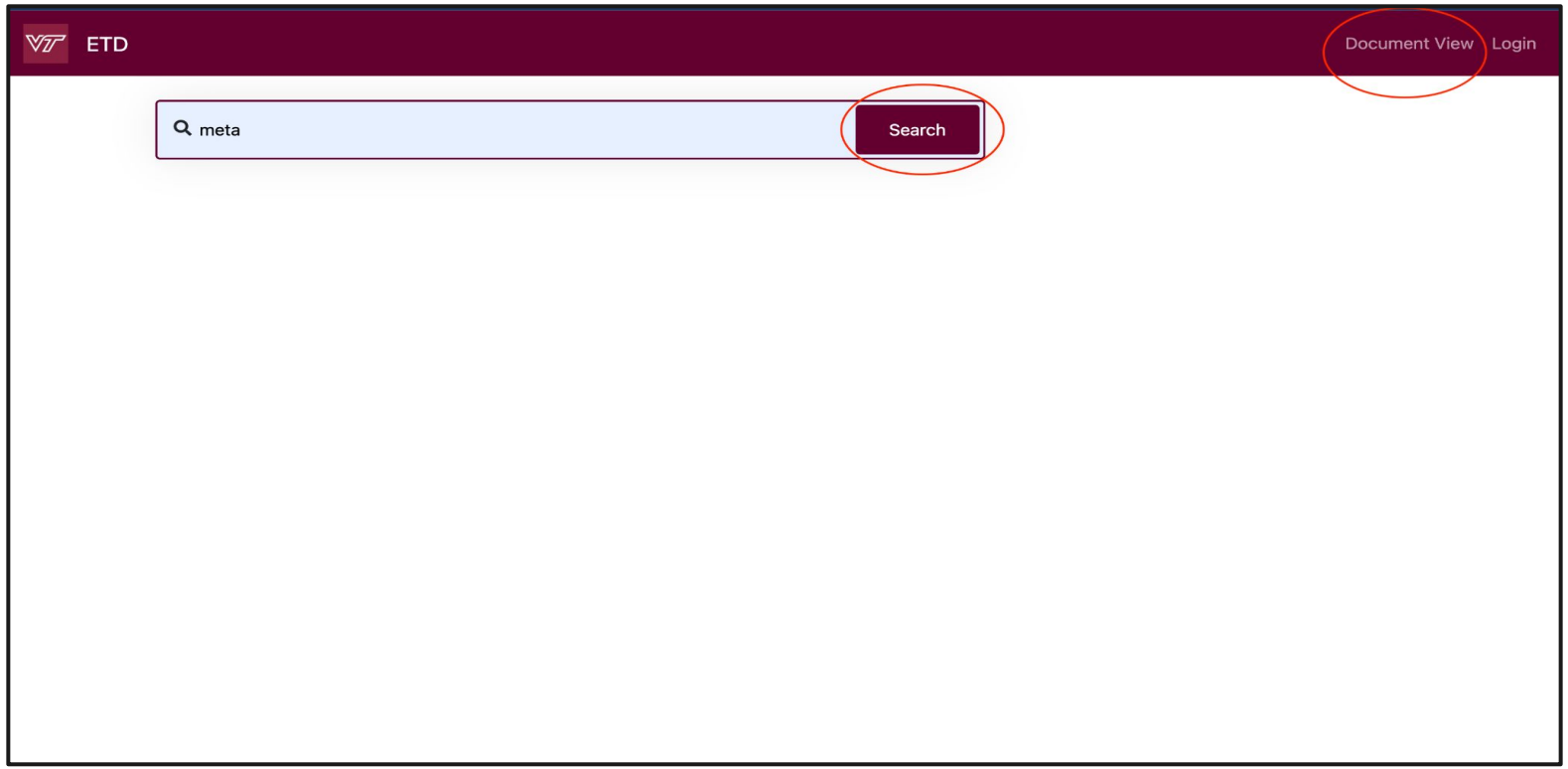
# Classification Model



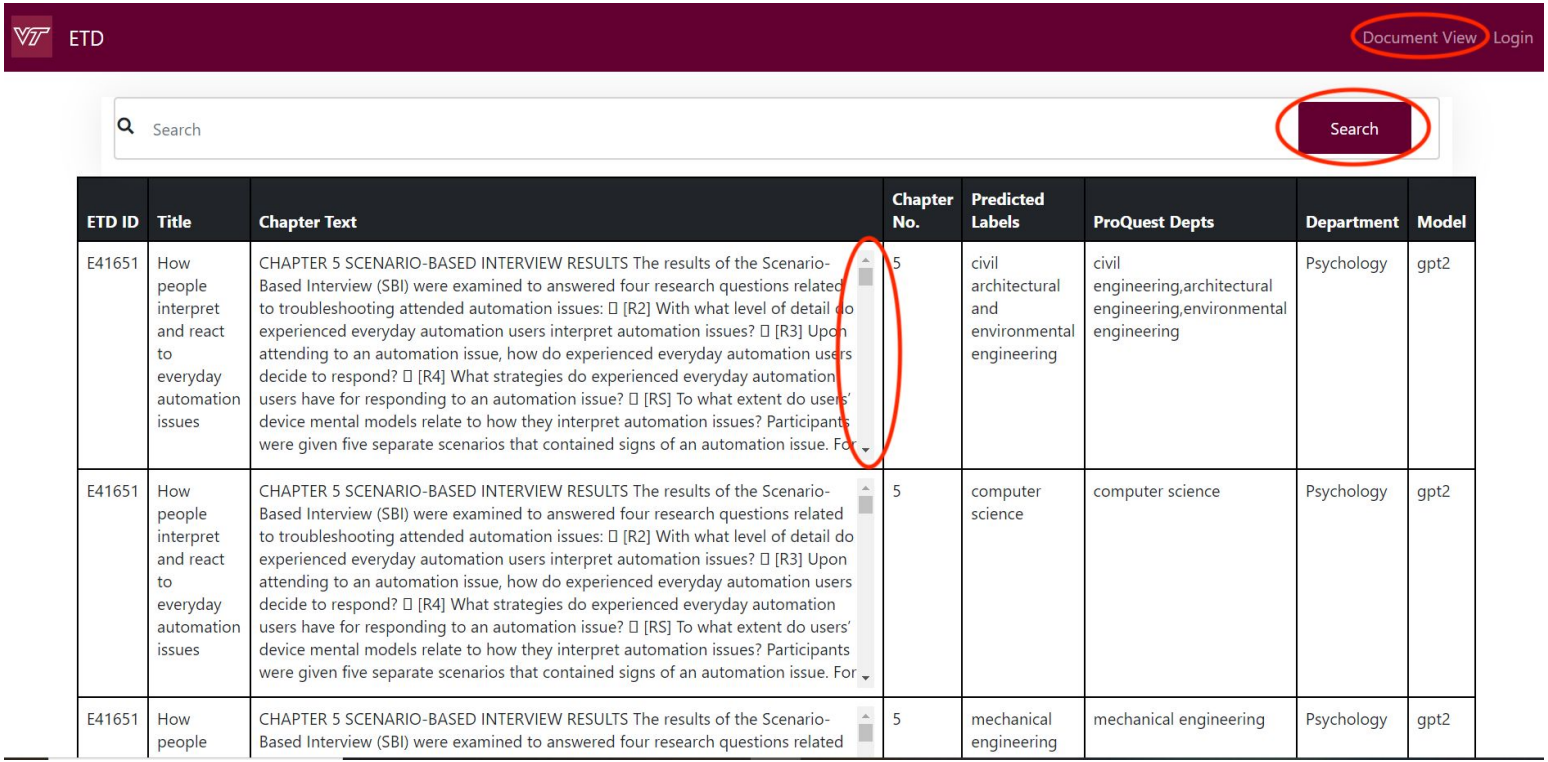
# Classification Model



# Document View Page



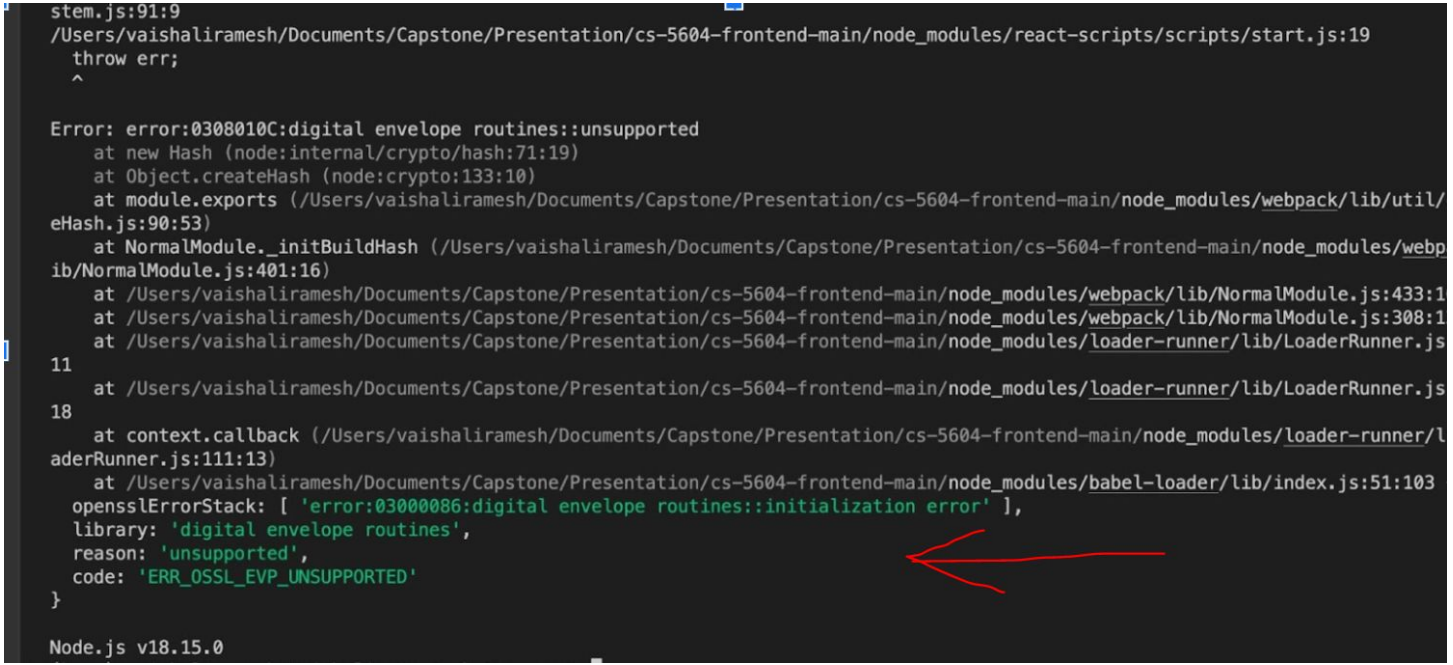
# Document View Page

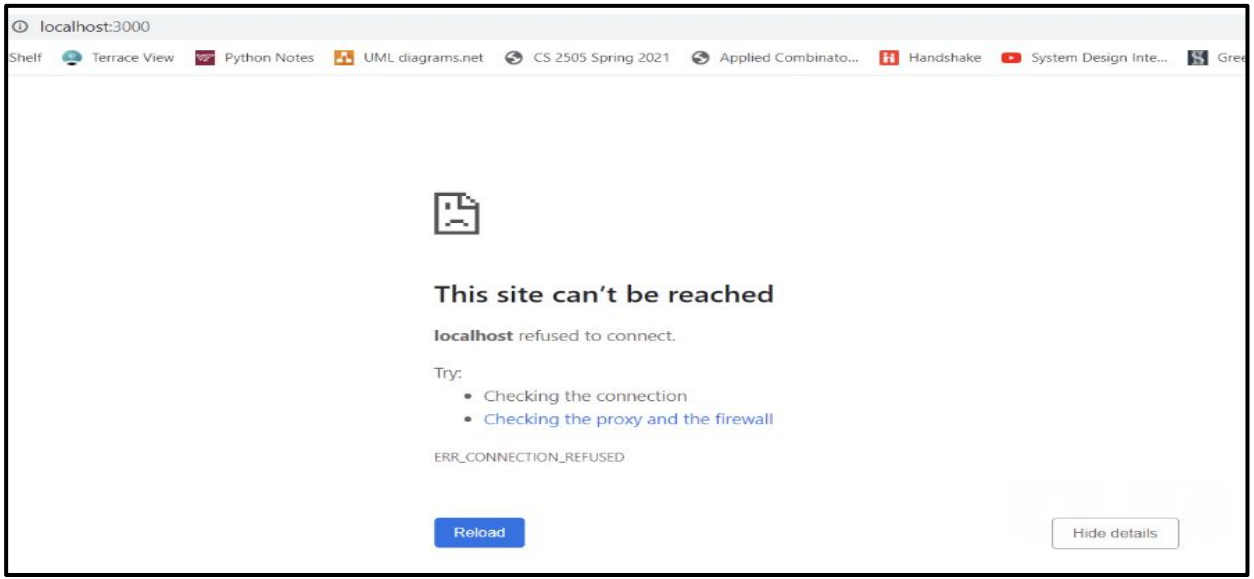



| ETD ID | Title | Chapter Text | Chapter No. | Predicted Labels | ProQuest Depts | Department | Model |
|---|---|---|---|---|---|---|---|
| E41651 | How people interpret and react to everyday automation issues | CHAPTER 5 SCENARIO-BASED INTERVIEW RESULTS The results of the Scenario-Based Interview (SBI) were examined to answered four research questions related to troubleshooting attended automation issues: □ [R2] With what level of detail do experienced everyday automation users interpret automation issues? □ [R3] Upon attending to an automation issue, how do experienced everyday automation users decide to respond? □ [R4] What strategies do experienced everyday automation users have for responding to an automation issue? □ [RS] To what extent do users' device mental models relate to how they interpret automation issues? Participants were given five separate scenarios that contained signs of an automation issue. For | 5 | civil architectural and environmental engineering | civil engineering,architectural engineering,environmental engineering | Psychology | gpt2 |
| E41651 | How people interpret and react to everyday automation issues | CHAPTER 5 SCENARIO-BASED INTERVIEW RESULTS The results of the Scenario-Based Interview (SBI) were examined to answered four research questions related to troubleshooting attended automation issues: □ [R2] With what level of detail do experienced everyday automation users interpret automation issues? □ [R3] Upon attending to an automation issue, how do experienced everyday automation users decide to respond? □ [R4] What strategies do experienced everyday automation users have for responding to an automation issue? □ [RS] To what extent do users' device mental models relate to how they interpret automation issues? Participants were given five separate scenarios that contained signs of an automation issue. For | 5 | computer science | computer science | Psychology | gpt2 |
| E41651 | How people | CHAPTER 5 SCENARIO-BASED INTERVIEW RESULTS The results of the Scenario-Based Interview (SBI) were examined to answered four research questions related | 5 | mechanical engineering | mechanical engineering | Psychology | gpt2 |

# Website Challenges

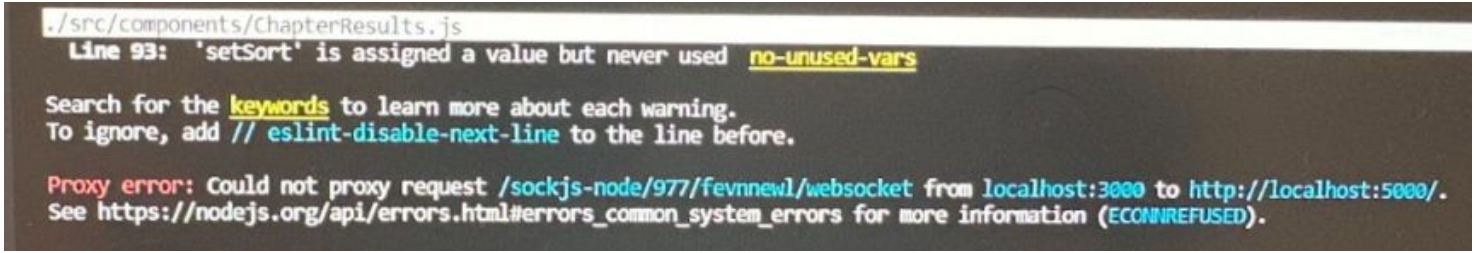
## Installation issues with Node.js and React.



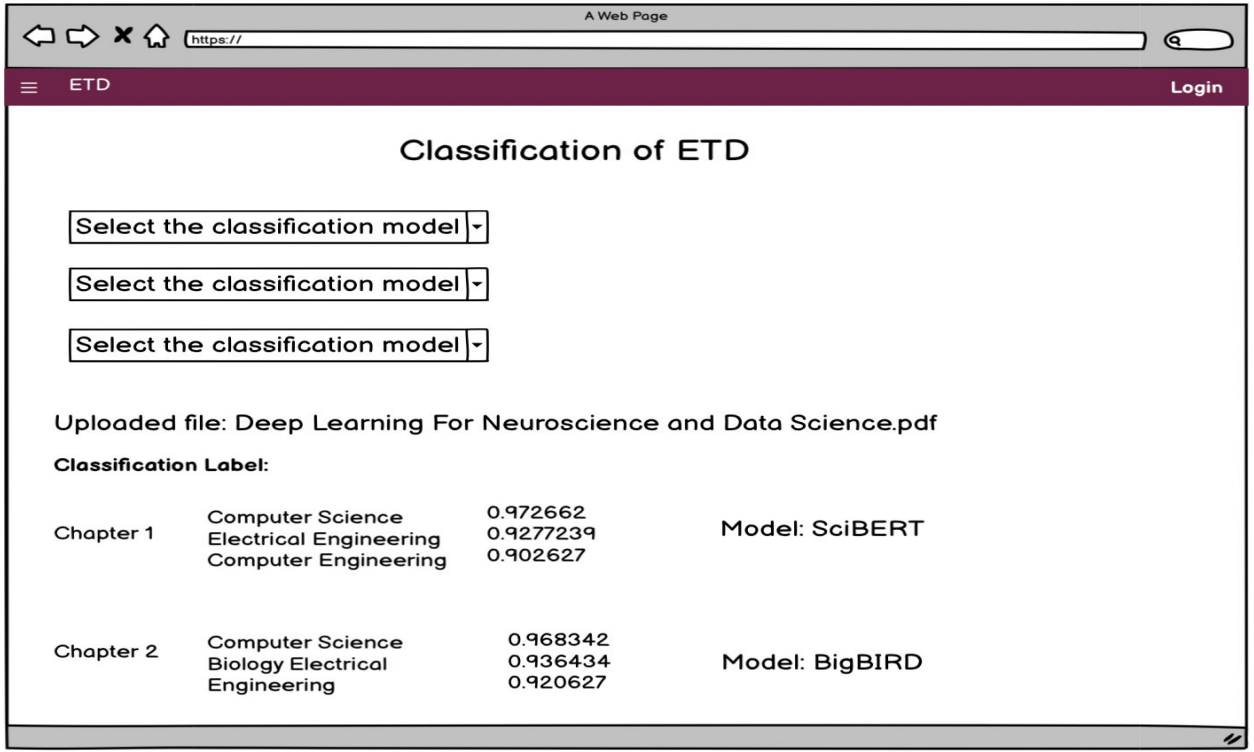



**Refusing connection to our local host**

**Proxy Error**



# Future Work - Website




- Hosting website on a Server
- Running Predictions from End-User POV.
- Testing the website

# Future Work - DL Models

- Analyze and Resolve embedding error for specific model and chapter combinations.
- Fine-Tune better Models for this task.

```
!python /content/run.py "bigbird" "/content/sample_data/chapter3_parsed.txt"

Text File Path is Valid, Proceeding ...
Using GPU: Tesla T4
Some weights of the model checkpoint at pszemraj/bigbird-pegasus-large-K-booksum were not used when initializing BigBirdPegasusModel: ['final_logits_bi
- This IS expected if you are initializing BigBirdPegasusModel from the checkpoint of a model trained on another task or with another architecture (e.g
- This IS NOT expected if you are initializing BigBirdPegasusModel from the checkpoint of a model that you expect to be exactly identical (initializing
../aten/src/ATen/native/cuda/Indexing.cu:1146: indexSelectLargeIndex: block: [42,0,0], thread: [32,0,0] Assertion `srcIndex < srcSelectDimSize` failed.
../aten/src/ATen/native/cuda/Indexing.cu:1146: indexSelectLargeIndex: block: [42,0,0], thread: [33,0,0] Assertion `srcIndex < srcSelectDimSize` failed.
../aten/src/ATen/native/cuda/Indexing.cu:1146: indexSelectLargeIndex: block: [42,0,0], thread: [34,0,0] Assertion `srcIndex < srcSelectDimSize` failed.
../aten/src/ATen/native/cuda/Indexing.cu:1146: indexSelectLargeIndex: block: [42,0,0], thread: [35,0,0] Assertion `srcIndex < srcSelectDimSize` failed.
```

```
!python /content/run.py "gpt2" "/content/sample_data/chapter3_parsed.txt"

Text File Path is Valid, Proceeding ...
Using GPU: Tesla T4
[['computer science', 'electrical and computer engineering', 'journalism']]
[[0.9925368428230286, 0.002030385425314307, 0.0007162086549215019]]
```

# Acknowledgements


- We would like to sincerely thank Dr. Edward A. Fox for the opportunity to work on this project, and for all his support and guidance throughout.
- We would also want to thank our client, Ms. Bipasha Banerjee for letting us work on this meaningful project, and for always helping and guiding us.
- Finally, we would like to acknowledge everyone who's currently working on this project or has worked on this project before us.

# References


Bethany Lord. 2021. What platform should I use to build my own website? (March 2021). Retrieved February 6, 2023 from https://theartsdevelopmentcompany.org.uk/resources/what-platform-should-i-use-to-build-my-own-website/

Gergana Petkova. 2022. The gold standard - the key to information extraction and data quality control. (May 2022). Retrieved February 6, 2023 from https://www.ontotext.com/blog/gold-standard-key-to-information-extration-data-quality-control/

Chanin Nantasenamat. 2021. How to build a machine learning model. (June 2021). Retrieved February 6, 2023 from https://towardsdatascience.com/how-to-build-a-machine-learning-model-439ab8fb3fb1

Sam Aenderson, Don Stroud, Dick Lyon, Karan Gupta. 2023. Outline of academic disciplines. (January 2023). Retrieved February 6, 2023 from https://en.wikipedia.org/wiki/Outline_of_academic_disciplines

Hugging Face Community Contributors (2019). Hugging Face-Pretrained Models - transformers 2.4.0 documentation. Retrieved from https://huggingface.co/transformers/v2.4.0/pretrained_models.html

# References

PyTorch Contributors. PYTORCH documentation(2023). Retrieved March 13, 2023 from https://pytorch.org/docs/stable/index.html

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., … Chintala, S. (2019, December 3). Pytorch: An imperative style, high-performance deep learning library. arXiv.org. Retrieved March 13, 2023 from https://arxiv.org/abs/1912.01703

CS5604 Fall 2022 Class. ETD project website.(2022). Retrieved March 13, 2023 from https://frontend.discovery.cs.vt.edu/

Javed Shaikh. (2017). "Machine Learning, NLP: Text classification using scikit-learn, Python, and NLTK." October 2017. Retrieved February 6, 2023 from https://towardsdatascience.com/machine-learning-nlp-text-classification-using-scikit-learn-python-and-nltk-c52b92a7c73a