**CSCE4290/5290 NATURAL LANGUAGE PROCESSING PROJECT IDEAS**

Below are some ideas for potential NLP-based projects that might help you and your team brainstorm on what you would want to work on for the term project:

# Text Classification

# **Research Paper Classification by Field.**

# Automatically categorize research papers based on academic fields.

# Example dataset: [arXiv Dataset](https://www.kaggle.com/datasets/Cornell-University/arxiv), [PubMed Articles](https://www.kaggle.com/datasets/owaiskhan9654/pubmed-multilabel-text-classification/data)

# **Medical Report Categorization**

# Classify medical reports into diagnoses, treatments, or patient history.

# Example dataset: [MIMIC-III Clinical Database](https://physionet.org/content/mimiciii/1.4/).

# **Multilingual Hate Speech Detection**

# Build a classifier to detect hate speech in multiple languages, addressing online toxicity and promoting responsible content moderation.

# Example dataset: [OLID](https://sites.google.com/site/offensevalsharedtask/olid), [hatespeech data](https://hatespeechdata.com/)

# **Code Summarization**

# Generate concise summaries of code snippets for better understanding and documentation.

# Example dataset: [CodeSearchNet Challenge dataset](https://ir-datasets.com/codesearchnet.html)

* Topic identification
  + Multi-label classification of printed media articles to topics
    - Dataset: [Greek Media monitoring multi-label classification](https://www.kaggle.com/c/wise-2014/data)
* Twitter Emotion Classification
* 6 different sentiments, with a multiclass classification problem
* Dataset: [Emotion Dataset for Emotion Recognition Tasks](https://www.kaggle.com/datasets/parulpandey/emotion-dataset)

# Text Summarization

# **Customer Review Summarization for Products**

# Summarize customer reviews for products into key insights

# Example dataset: [Amazon Product Reviews](https://www.kaggle.com/datasets/saurav9786/amazon-product-reviews)

# **Multimodal Summarization for Scientific Papers**

# Combine text analysis of scientific papers with image and caption analysis to generate concise summaries highlighting key figures, results, and conclusions.

# Example dataset: [S2ORC](https://github.com/allenai/s2orc?tab=readme-ov-file#download-instructions), [CORD-19](https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge)

* Automatically create a summary with the major points of the original document?
  + Abstractive (write your own summary) and Extractive (select pieces of text from original) are two popular approaches
    - Dataset: [CNN and DailyMail News Pieces](http://cs.nyu.edu/~kcho/DMQA/) by Google DeepMind

# Sentiment Analysis

# **Sentiment Analysis in Social Media Live Streams**

# Analyze sentiment in real-time during live streams like sports games or political debates, considering speaker dynamics and audience reactions.

# Dataset: [Twitter Open Sentiment Challenge](https://www.kaggle.com/code/tarunpaparaju/twitter-challenge-roberta-sentiment-predictor/input)

# **Multilingual Sentiment Analysis for Global Reach**

# Develop a sentiment analysis model that can handle multiple languages effectively to analyze opinions and feedback from diverse audiences.

# Dataset: [Amazon Reviews Multilingual](https://www.kaggle.com/datasets/mexwell/amazon-reviews-multi)

* Generic Sentiment Analysis
  + Tweets sorted by geography and timestamp.
    - Dataset: [Tweets sentiment tagged by humans](https://inclass.kaggle.com/c/si650winter11/data)
  + Movie reviews
    - Datasets such as [Movie Review Data](https://www.cs.cornell.edu/people/pabo/movie-review-data/) and others
  + Amazon product reviews
* Aspect-based sentiment analysis
* Identify specific aspects of products (e.g., price, features, performance) mentioned in reviews and analyze sentiment towards each aspect.
* Dataset: [Yelp Reviews with Ratings and Aspects](https://www.yelp.com/dataset)

# Speech Recognition/Conversational AI

* Social Chat/Conversational Bots
  + Can you build a bot which talks to you just like people talk on social networking sites?
    - Reference: [Chat-bot architecture](https://github.com/aryanc55/TS3000_TheChatBOT)
    - Dataset: [Reddit Dataset](http://files.pushshift.io/reddit/comments/)
* Copy-cat Bot
  + Generate plausible new text which looks like some other text
  + Obama Speeches? For instance, you can create a bot which writes some [new](https://medium.com/%40samim/obama-rnn-machine-generated-political-speeches-c8abd18a2ea0) [speeches in Obama's style](https://medium.com/%40samim/obama-rnn-machine-generated-political-speeches-c8abd18a2ea0)
  + Trump Bot? Or a Twitter bot which mimics [@realDonaldTrump](http://www.twitter.com/%40realdonaldtrump)

# Language Understanding

# Multilingual Question Answering

# predicting the answers to questions that are asked in different languages.

# Dataset : [Hindi and Tamil Question Answering](https://www.kaggle.com/competitions/chaii-hindi-and-tamil-question-answering/data)

# **Cybersecurity Threat Detection from Multilingual Text Data**

# Analyze text-based cyber threats like phishing emails or malware alerts across languages, with explainable reasoning for improved proactive security measures.

# Dataset: [OLID](https://sites.google.com/site/offensevalsharedtask/olid), [hatespeech data](https://hatespeechdata.com/)

* Automated essay grading
  + The purpose of this project is to implement and train machine learning algorithms to automatically assess and grade essay responses.
    - Dataset: [Essays with human graded scores](https://www.kaggle.com/c/asap-aes/data)
* Sentence to Sentence semantic similarity
  + Can you identify question pairs that have the same intent or meaning?
    - Dataset: [Quora question pairs](https://www.kaggle.com/c/quora-question-pairs/data) with similar questions marked
* Fight online abuse
  + Can you confidently and accurately tell whether a particular comment is abusive?
    - Dataset: [Toxic comments on Kaggle](https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge)
* Open Domain question answering
  + Can you build a bot which answers questions according to the student's age or her curriculum?
  + [Facebook's FAIR](https://github.com/facebookresearch/DrQA) is built in a similar way for Wikipedia.
    - Dataset: [NCERT books](http://ncert.nic.in/ebooks.html) for K-12/school students in India, [NarrativeQA by Google DeepMind](https://github.com/deepmind/narrativeqa) and [SQuAD by Stanford](https://rajpurkar.github.io/SQuAD-explorer/)

Other ideas include:

# Recommender systems (based on user's social media profiles)

* **Machine Translation**
* **Speech Recognition**

**Integration of Two Topics:**

For example, consider two topics A, B

**Method 1:** Work on topic A and the output of A should be the input of Topic B.

Examples:

1. **Machine Translation + Question Answering**

Develop a system that translates claims made in English to Hindi and Tamil, analyzes their factual accuracy based on relevant evidence sources, and answers the questions asked in each translated language.

1. **Machine Translation + Text Summarization**

Translate documents from one language to another and create a model to generate concise summaries of the translated documents in the target language.

1. **Question Answering + Contextual Summarization**

A model that retrieves relevant documents answering a user's question, then automatically generates a concise summary highlighting the key information for a complete understanding.

1. **Language Understanding + Question Answering:**

Train a language understanding model to extract meaning from open-ended questions (topic a), then leverage that understanding to answer the questions using a dedicated QA system (topic b).

**Method 2:** If you consider two simple topics**.** Work on topic A and fine tune the model of topic A and work on Topic B and fine tune the model of topic B.

Example:

1. **Text Classification (Fine-tuned) + b) Summarization (Fine-tuned)**

**Topic A:** Pre-trained language model (e.g., BERT) on a specific classification dataset related to your project's focus (e.g., scientific articles by domain).

**Topic B:** Pre-trained summarization model (e.g., T5) on summaries conditioned on the predicted class from topic A (e.g., summaries specific to each scientific domain).

**Project Goal:** Generate domain-specific summaries for classified documents, improving information access and understanding.

**Method 3:** Work on Topic A and Topic B and then fine tune the model

Example:

1. **Speech Recognition + Text Summarization (Fine – tuned)**

Combining speech recognition with text summarization and fine-tuning the integrated model. This combination would involve first transforming spoken language into text using speech recognition and then summarizing this text to extract key points or themes. The summarization component would need to be fine-tuned to effectively condense the transcribed text, ensuring that the summaries are coherent, relevant, and capture the essence of the spoken content.

**References:**

<https://github.com/NirantK/awesome-project-ideas#text> [Machine Learning Mastery Blog](http://machinelearningmastery.com/text-generation-lstm-recurrent-neural-networks-python-keras/)

[Word Embeddings vs. Words](https://www.kaggle.com/c/word2vec-nlp-tutorial#description)