Analysis

This analysis involved extracting data from Reddit over two years, manually labeling posts into specific categories, training machine learning models, and performing exploratory data analysis (EDA) and topic modeling to glean insights into AI and employment discussions.

# Data Collection and Preprocessing

## Data Extraction

Due to limitations with Reddit’s API, we utilized the Reddit data archive, which is accessible via torrents. The data spans from 2005 to 2023, but we focused on posts from July 2022 to July 2024. The data was downloaded using academic torrents and processed using scripts from the [PushshiftDumps repository](https://github.com/Watchful1/PushshiftDumps).

**Data Statistics:**

* Total Reddit posts matching the query from the ten subreddits between July 2022 and July 2024: **7,616,585**.
* The number of Reddit posts labeled by Llama 3.1 7B (LLM) was around **18,159**.

**Data Labeling**

We manually labeled **555** Reddit posts using multilabel classification into three categories (levels):

1. **C1 Work**: Posts discussing task-oriented thoughts, such as new summarization models.
2. **C2 Worker**: Posts about how a worker might have suffered impact their work environment.
3. **C3 Workforce**: Posts indicating that a significant number of workers in a specific sector or in general are affected.

**Label Distribution:**

|  |  |
| --- | --- |
| **Category** | **Number of Posts** |
| C1 Work | 67 |
| C2 Worker | 192 |
| C3 Workforce | 80 |

# Methodology

## Exploratory Data Analysis (EDA)

EDA was conducted on the manually labeled and predicted data to understand trends and patterns.

1. **Time Series Analysis**: Plots showing the number of Reddit posts per month were created to identify trends over time, with a special focus on significant events like the release of ChatGPT on November 30, 2022.

A graph with blue lines

Description automatically generated

A graph with blue lines

Description automatically generated

A graph of a graph

Description automatically generated

A graph with a line

Description automatically generated

A graph of different colored lines

Description automatically generated

1. **Sentiment Analysis**: The average sentiment per month was calculated using TextBlob’s sentiment analyzer to assess the emotional tone of the discussions.

A graph of blue lines

Description automatically generatedA graph of a graph

Description automatically generated with medium confidenceA graph with blue lines

Description automatically generated

A graph with a line

Description automatically generated

A graph of different colored lines

Description automatically generated

# Machine Learning Models

We trained two machine learning models to predict the categories of unlabeled data:

1. Random Forest Classifier
2. Support Vector Machine (SVM)

## Data Preprocessing:

* Text data was cleaned by converting it to lowercase and removing URLs, punctuation, numbers, and stop words.
* Lemmatization was applied to reduce words to their base forms.
* The cleaned text was transformed into numerical features using TF-IDF vectorization.

## Model Training:

* The models were trained using the manually labeled data through multilabel classification (the categories are not mutually exclusive)
* The data was split into training and validation sets using an 80/20 split.
* We used the following evaluation metrics: Precision, Recall, F1-Score, and Support (number of occurrences of each class in the validation set)

**Model Performance:**

**Random Forest:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Category** | **Precision** | **Recall** | **F1-Score** | **Support** |
| C1 Work | 1.00 | 0.10 | 0.18 | 10 |
| C2 Worker | 0.61 | 0.33 | 0.43 | 33 |
| C3 Workforce | 0.00 | 0.00 | 0.00 | 16 |

**SVM:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Category** | **Precision** | **Recall** | **F1-Score** | **Support** |
| C1 Work | 0.00 | 0.00 | 0.00 | 10 |
| C2 Worker | 0.69 | 0.27 | 0.39 | 33 |
| C3 Workforce | 0.50 | 0.12 | 0.20 | 16 |

**Predicted Posts per Category and Model:**

|  |  |  |
| --- | --- | --- |
| **Category** | **Random Forest** | **SVM** |
| C1 Work | 0 | 10 |
| C2 Worker | 920 | 942 |
| C3 Workforce | 7 | 61 |

**Model Selection:**

* Both models struggled with predicting C1 Work and C3 Workforce due to low support and class imbalance.
* The Random Forest model had a higher F1-Score for C2 Worker, so it was selected for further analysis.

# Topic Modeling

Topic modeling was performed using Latent Dirichlet Allocation (LDA) on four datasets:

1. Manually Labeled Data for All 3 Categories:
   1. C1 Work
   2. C2 Worker
   3. C3 Workforce
2. Predicted Data for C2 Worker

**Results Interpretation:**

* The topics generated were interpreted to understand prevalent themes in each category.

**A graph with a bar chart and a bar chart with a bar chart and a bar chart with a bar chart and a bar chart with a bar chart and a bar chart with a bar chart and

Description automatically generated**

*C1 Work LDA*

*A screenshot of a graph

Description automatically generated*

*C2 Worker LDA*

*A graph with text and numbers

Description automatically generated with medium confidence*

*C3 Workforce LDA*

*A graph with blue circles and black text

Description automatically generated*

*Predicted C2 Worker LDA*

**Key Findings:**

* **C1 Work:** Discussions focused on integrating AI into specific tasks and the potential displacement or transformation of jobs.
* **C2 Worker:** Personal and professional concerns about AI’s impact on individual workers, especially in creative fields like art.
* **C3 Workforce:** Systemic concerns about AI’s impact on the workforce include job displacement and the rise of Artificial General Intelligence (AGI).
* **Predicted C2 Worker:** Reflected a mix of personal concerns about AI’s impact on careers and broader discussions about technological advancements.

# Summary

**EDA Findings**

* Trends Over Time: There was an observable increase in discussions on AI and employment following the release of ChatGPT.
* Sentiment Analysis: The average sentiment fluctuated over time, indicating varying emotional responses to AI developments.
* Word Frequencies: Common words included “ai,” “job,” “people,” “work,” and “human,” highlighting the central themes of employment and human-AI interaction.

**Model Performance**

* Random Forest Model: Demonstrated better performance for the C2 Worker category but failed to effectively predict C1 Work and C3 Workforce.
* SVM Model: Showed limited effectiveness across all categories.
* Challenges: The models struggled due to class imbalance and the complexity of the multilabel classification task.

**Topic Modeling Insights**

* Common Themes: AI’s impact on employment, human vs. AI creativity, and future uncertainties were prevalent across all categories.
* Unique Aspects in Predicted Data: Technical discussions about AI development and ethical considerations were included but not emphasized in the manually labeled data.
* Worker Concerns: Reflected anxiety, curiosity, and skepticism regarding AI, indicating the emotional impact on individuals.