Analysing AI Related Job Anxiety and Future Optimism via Text Data

IST 332: Final Project Report

Aashish Sunar, Dolma Rawat, Rohini Vishwanathan & Yashas Basavaraju Mahesh

GitHub Repository: https://github.com/dolmarawat/NLP-332-Final-Project.git
Google Colab Link: https://colab.research.google.com/drive/1OPeNnSsnKgoP0D-c6eSSt-f0VXiqC4j_?usp=sharing

Abstract

Artificial intelligence is reshaping not only labor markets but also how individuals perceive their career futures, job security, and professional identities. While existing research has largely focused on estimating automation risk, task displacement, and labor market impacts, less attention has been paid to the emotional dimensions of technological change as expressed in everyday discourse. This study examines how emotional orientations toward AI and work are constructed linguistically in online communities.

Using a corpus of approximately 10,000 Reddit posts collected from AI- and career-focused subreddits, the study applies a multi-stage natural language processing pipeline combining sentiment analysis, topic modeling, and supervised classification. Sentiment analysis and a hybrid rule-based labeling framework identify three dominant emotional categories: **Career Anxiety**, **Future Optimism**, and **Uncertainty**. Topic modeling using classical and embedding-based methods reveals that anxiety-oriented discourse centers on layoffs, burnout, and market instability, while optimistic discourse emphasizes innovation, skill development, and career mobility. Uncertainty emerges as a distinct and meaningful emotional stance characterized by exploratory and mixed affect.

Supervised learning experiments demonstrate that emotional orientation is predictively learnable from text, with linear models—particularly Support Vector Machines—achieving the strongest performance on TF-IDF representations. Model comparison highlights the importance of aligning feature representation with classifier choice rather than relying on model complexity alone.

By integrating computational methods with socio-emotional analysis, this study provides insight into how individuals emotionally negotiate technological change in real time. The findings contribute to computational social science and AI governance research by foregrounding emotional experience as a critical dimension of labor market transformation.

Keywords: artificial intelligence, labor markets, emotion analysis, natural language processing, sentiment analysis, topic modeling, Reddit

Table of Contents

1. Introduction

1.1 Background & Motivation

Artificial intelligence is not only reshaping labor markets but also restructuring how individuals perceive their own economic futures, career trajectories, and professional identities. While traditional economic analyses have primarily focused on quantifying automation risk, job displacement probabilities, and skill-task complementarities (Autor, 2015; Frey & Osborne, 2017; Acemoglu & Restrepo, 2019), these approaches often miss the lived emotional realities that accompany rapid technological change. For many workers, AI is not experienced as an abstract macroeconomic force but as an immediate psychological and professional pressure—manifesting as fears of redundancy, uncertainty about career stability, or anxiety about reskilling expectations.

At the same time, AI has cultivated a parallel discourse of excitement and possibility. Communities enthralled by Artifical General Intelligence breakthroughs, multimodal models, and exponential innovation often express deep optimism, technological enthusiasm, and belief in a dramatically transformed future. Thus, public narratives around AI occupy a fragmented emotional spectrum: **AI-related job anxiety** coexists with **AI-driven future optimism**, often within the same online ecosystems.

Reddit, one of the largest public forums for peer-to-peer knowledge exchange, has become a crucial site for observing these emotional and cognitive negotiations. Workers use subreddits such as r/cscareerquestions, r/techlayoffs, and r/artificial as spaces for vulnerability, frustrated reflection, and real-time storytelling. They share experiences of layoffs, toxic workplaces, burnout, and unstable hiring cycles—narratives that economists typically cannot capture in quantitative models. Other communities, such as r/Singularity or r/OpenAI, construct bold visions of a technological future marked by discovery, progress, and transcendence.

These discussions provide an invaluable opportunity to analyze how emotional orientations toward AI are linguistically constructed in everyday discourse. They offer clues about emerging labor fears, public reactions to innovation, and the psychological landscape shaping technological acceptance, resistance, or adaptation. Yet despite their richness, such online narratives remain understudied within computational social science, NLP, and technology governance research.

1.2 Research Problem

Although research in economics, sociology, and human–computer interaction has examined how automation influences labor markets, productivity, and job design, much less attention has been given to the *ways individuals talk about these changes in real time*. Emotional responses—especially those tied to lived experiences of job insecurity or excitement about technological futures—remain largely unmodeled in the NLP space.

We recognized that public discussions about AI often collapse into general "positive vs. negative sentiment," but the conversations we observed were more layered. Posts about layoffs were not always purely negative; some included resilience, relief, or reflection. Optimistic posts were not always purely positive; some contained caution or uncertainty. This complexity motivated us to explore whether more subtle emotional distinctions, such as **AI-related job anxiety** and **AI-related future optimism**, could be described and predicted through linguistic features.

The core research problem we address is therefore twofold:

1. **Understanding how people linguistically construct their anxieties and hopes around AI**, particularly in relation to work.
2. **Determining whether these emotional orientations can be modeled computationally** using sentiment analysis, topic modeling, and supervised learning

1.3 Research Objectives

In shaping this project, we set out to understand how people describe their experiences with AI in relation to work—and whether these emotional perspectives can be analyzed and modeled from language alone. To guide our investigation, we defined a set of objectives that reflect both the exploratory and predictive dimensions of our work.

1. **Construct a meaningful dataset that reflects real conversations about AI and work:** We aimed to build a corpus that captures the authentic language people use when discussing job security, layoffs, career transitions, technological optimism, and future expectations. The goal was not just to gather text, but to collect discourse that represents the emotional and thematic diversity of the topic.
2. **Prepare and refine the text so the underlying linguistic patterns become visible:** High-quality analysis requires clean and consistent data. Through systematic preprocessing—normalizing text, lemmatizing, removing noise, and examining word behavior—we sought to reveal the structural and semantic characteristics that shape AI-related discourse.
3. **Understand how emotions appear in language and identify cues that distinguish anxiety from optimism:** By analyzing sentiment distributions, contextual word usage, and variation in emotional intensity, we aimed to uncover how people signal stress, fear, hope, and enthusiasm in their writing. This step was central to understanding whether emotional orientation could be detected computationally.
4. **Identify the major themes that organize discussions about AI and work:** Beyond individual posts, we were interested in the broader topics that consistently emerge across conversations—whether they relate to layoffs, reskilling pressures, workplace culture, AGI speculation, or long-term future visions. Topic modeling allowed us to map these shared themes and compare how they differ across emotional frames.
5. **Build and evaluate models that can recognize emotional orientation from text:** A key question in our project was whether the distinction between anxiety-oriented and optimism-oriented posts is learnable. By training supervised models, we aimed to test how well linguistic features, embeddings, and textual patterns support classification and sentiment prediction.
6. **Translate our findings into insights that matter for broader conversations about AI and work:** Finally, we wanted our analysis to be meaningful beyond the technical workflow. Our goal was to highlight how people interpret AI-related changes, to identify patterns that may help institutions anticipate emerging concerns, and to contribute a grounded perspective on how public narratives shape responses to technological transformation.

1.4 Expected Impact

Through this project, we aim to contribute a clearer picture of how individuals navigate the uncertainty and possibility brought about by AI. Our findings have potential value across multiple domains:

- **Workforce and HR analytics:** Insights into public expressions of stress, insecurity, or optimism can inform decisions related to employee support, reskilling initiatives, and communication strategies.

- **Education and career development:** Understanding how students and professionals articulate AI-related concerns may help institutions tailor guidance, advising, and curriculum design.
- **AI governance and public communication:** Policymakers and organizations benefit from understanding how narratives around AI reshape trust, risk perception, and expectations.
- **Computational research:** Our work integrates corpus creation, linguistic analysis, sentiment modeling, topic modeling, and supervised learning, offering a structured pipeline for analyzing emotionally complex online discourse.

Ultimately, this project helps reveal how people understand and emotionally respond to AI—not in theory, but in their own words. By examining these conversations, we hope to support efforts that address the anxieties people face while also recognizing the optimism and curiosity that continue to drive technological engagement.

2. Corpus Creation

Corpus creation for this project focused on building a 10,000-item Reddit text dataset about AI, software careers, and job insecurity, using a custom Python scraper and saving the final corpus as reddit_reviews_10k_FINAL.csv . The goal was to capture both posts and comments from highly relevant subreddits so that the downstream NLP models could analyze real-world attitudes, fears, and experiences around AI and tech careers. This section describes the data sources, collection pipeline, and corpus statistics in a way that is suitable to paste directly into the report.

2.1 Data Sources

The corpus was constructed entirely from public Reddit content, targeting subreddits where people discuss AI, software jobs, layoffs, and future-of-work topics. The scraper used a curated dictionary of subreddits mapped to high-level thematic labels, such as :

- cscareerquestions and Layoffs -> CAREERANXIETY (software careers, layoffs, work conditions)
- Singularity, Futurology and OpenAI -> FUTUREHYPE (hype, optimism, fear, and speculation about AI and the future)

Each text sample in the final CSV carries the corresponding label so different discourses (career anxiety vs. AI futurism) can be compared. All data was collected from the "top" posts of each subreddit (top posts over the last month, limited to 100 per subreddit) to ensure reasonably high-quality, discussion-rich threads.

2.2 Collection Methods

The corpus was created using a custom Python script, scraper.py, written and then executed. The script relied on the Reddit JSON endpoints rather than any unofficial HTML scraping to keep the pipeline simpler and more robust.

- Target size: TARGET_COUNT = 10,000, i.e., the script stops once at least 10,000 text items are collected.
- Maximum comments per post: COMMENTS_PER_POST = 60, to avoid over-sampling a single thread while still capturing rich discussions.
- Checkpoint interval: CHECKPOINT_INTERVAL = 500, with automatic partial saves to reddit_reviews_partial.csv every 500 items as a safety net against crashes or rate-limit failures.
- User-Agent header: a realistic browser user agent string was set in the headers dictionary to reduce the chance of being blocked by Reddit.

The script iterated over the predefined subreddit label dictionary, and for each subreddit it:

- Created a dedicated listing URL to programmatically fetch the platform's top-performing posts, serving as the critical data foundation for analysis and display.
- Parsed the JSON response, extracted the list of posts, and then looped through each post's permalink.
- For each permalink, a helper function to scrape the main post and its comments.

2.2.1 HTTP and rate-limit handling

To make the scraper reasonably robust, two helper functions were used :

- get_json(url) : attempts up to three HTTP GET requests with a timeout, handles status code 429 (Too Many Requests) by sleeping, and returns parsed JSON or None on repeated failure.
- scrape_post_and_comments(permalink, label) : builds the .json URL for the thread, fetches it via get_json, and then extracts the post text and comments.

The script adds short time.sleep pauses between requests (for example, around 1.5 seconds in the main loop) to be gentle on Reddit's servers and reduce rate limiting. When Reddit returns 429 or a network error occurs, the code waits and retries, which allows the crawl to progress steadily without manual intervention.

2.2.2 Text extraction and filtering

Within each thread, the scraper processes both the root post and the top-level comments:

- Post body : the script pulls title and selftext from the post JSON, concatenates them, and filters out very short selftexts (e.g., requiring at least 50 characters) to avoid low-information posts.
- Comment body: it iterates over the data[1].data.children array, treating each element as a potential comment and reading the body field.

Before appending any text to the dataset, several filtering rules are applied :

- Skip deleted comments: comments whose body is "[deleted]" or otherwise missing are ignored.
- Minimum length: comments shorter than about 30 characters are skipped to avoid noise such as "lol" or "same".
- Comment cap: no more than COMMENTS_PER_POST comments are collected per thread, even if more are available.

For every accepted piece of text, the script appends a record of the form :

- **Text** : full text content (title + selftext for posts, or the comment body for comments).
- **Label** : The high-level category derived from the subreddit (e.g., CAREERANXIETY, FUTUREHYPE).
- **Type** : a simple indicator of whether the row came from a post or a comment.

Throughout the run, progress messages are printed to the console (for example, indicating which subreddit is being mined, how many posts are in the listing, and the current total count).

2.2.3 Checkpoints and Final save

To protect against losing progress, the dataset is periodically converted to a pandas DataFrame and saved :

- Every time the total number of collected items crosses a multiple of CHECKPOINT_INTERVAL (500), the script calls save_checkpoint(all_reviews) and writes reddit_reviews_partial.csv to disk.
- After the main loop over all subreddits finishes or the target count is reached, the script creates a final DataFrame and saves it as reddit_reviews_10k_FINAL.csv with index=False.

A final log line prints the total number of rows in the saved file, confirming that the corpus size requirement is met.

2.3 Corpus Statistics

The final corpus file, reddit_reviews_10k_FINAL.csv, contains just over ten thousand rows of Reddit text samples, each with its associated label and type. Rows are stored in plain text form, with examples showing multi-sentence comments and posts that mix narrative, opinion, and discussion about layoffs, AI fears, hiring processes, remote work, and related topics.

At a high level, the CSV includes:

- A text-like column with the full Reddit content including both short and very long entries, depending on the original comment or post.
- A label column that captures subreddit-level themes (for example, CAREERANXIETY or FUTUREHYPE), which can be used as the target variable or as a grouping variable for analysis.
- A type column marking whether the entry is a post or comment, allowing separate analyses of original posts versus replies.

Many entries are fairly long, reflecting in-depth discussions; for instance, some rows span multiple paragraphs about software engineers' workload, hiring frustrations, fears of AI-driven layoffs, or macroeconomic concerns around AI investment. This richness makes the corpus suitable for tasks like topic modeling, sentiment analysis, stance detection, or comparing discourse patterns between career-focused and future-focused communities.

3. Text Preprocessing

3.1 Overview

Preprocessing was conducted to standardize the heterogeneous Reddit text and prepare it for linguistic, sentiment, and supervised learning analyses. User-generated content on social platforms typically contains informal constructions, variable orthography, punctuation artifacts, emojis, contractions, and platform-specific tokens. Without normalization, such noise reduces the interpretability and stability of downstream models. Accordingly, a comprehensive preprocessing pipeline was implemented to ensure that the corpus was analytically tractable while preserving semantically meaningful content.

3.2 Preprocessing Pipeline

The preprocessing workflow consisted of a multi-stage sequence of operations integrating both NLTK and spaCy libraries. The steps below reflect the exact order and procedures applied to the full corpus.

- **Contraction Expansion:** Reddit posts frequently include contractions, which complicate token standardization. The *contractions* package was applied to expand contracted forms (e.g., *don't → do not*, *I'm → I am*), improving lexical consistency prior to tokenization.
- **Lowercasing:** All text was converted to lowercase to eliminate case-based duplication and maintain uniform vocabulary representation.
- **Removal of Digits and Punctuation:** Numeric tokens, punctuation marks, and tokens consisting exclusively of punctuation were removed using NLTK utilities and regular expression filters. This step reduced non-semantic noise without influencing the content-bearing structure of the text.
- **Tokenization:** The NLTK tokenizer segmented each document into a sequence of lexical units, forming the basis for subsequent linguistic operations.
- **Stopword Removal:** Stopwords were removed using the standard NLTK English stopword list, supplemented with a custom extension designed to eliminate Reddit-specific artifacts. This ensured that analytical attention remained on semantically relevant expressions.
- **Length-Based Filtering:** Tokens with fewer than three characters were removed, with the exception of meaningful domain-specific tokens such as *ai*. This step mitigated noise associated with short fillers and internet shorthand.
- **Lemmatization (spaCy):** Lemmatization was performed using spaCy's *en_core_web_sm* model, reducing words to their dictionary base forms (e.g., *workers → worker*, *studying → study*). Lemmatization reduced sparsity, enhanced interpretability, and improved the alignment of tokens across grammatical variations.
- **Stemming (Snowball Stemmer):** To further reduce morphological variation, stemming was applied using the Snowball Stemmer. This step was particularly useful for feature extraction in TF–IDF and bag-of-words representations. Words such as *management*, *manager*, and *managing* were simplified to the root *manag*, improving feature compactness for supervised learning.

3.3 Illustrative Example

To document the transformation of raw Reddit text into structured and model-ready tokens, three representative snapshots from the preprocessing pipeline are presented below.

(a) Raw Corpus Structure The initial dataset consisted of three fields—*text*, *label*, and *type*—as shown in Figure 3.1. Each entry was categorized as either *Career Anxiety* or *Future Optimism*, based on manual review of thematic and emotional cues.

Figure 3.1. Sample of the raw corpus prior to preprocessing

```
import pandas as pd
df = pd.read_csv('/content/Drive/MyDrive/IST332_NLP/reddit_reviews_10k_FINAL.csv')
df.head()
```

	text	label	type
0	I GOT THE JOB!! F*** MY OLD MANAGER!!! I've ha...	CAREER_ANXIETY	post
1	congrats! just a word of caution not to tell t...	CAREER_ANXIETY	comment
2	Congrats on the pay raise! Did you study anyth...	CAREER_ANXIETY	comment
3	update to us how you break the news to your ma...	CAREER_ANXIETY	comment
4	Congrats. You can do the bare minimum now and ...	CAREER_ANXIETY	comment

This structure provided the foundation for subsequent linguistic normalization and label-based analysis.

(b) Intermediate Stemming Output

Following tokenization, stopword removal, punctuation stripping, and stemming, each document was represented as a sequence of morphological root forms. As illustrated in Figure 3.2, this step consolidated related variants (e.g., *managing, manager, management → manag*), reducing sparsity in the feature space.

Figure 3.2. Intermediate representation showing stemmed tokens

	Review_Stemmed
0	[i, got, the, job, !, !, f, *, *, *, my, old, ...
1	[congrat, !, just, a, word, of, caution, not, ...
2	[congrat, on, the, pay, rais, !, did, you, stu...
3	[updat, to, us, how, you, break, the, news, to...
4	[congrat, ., you, can, do, the, bare, minimum,...
5	[hope, you, thrive, at, this, one, and, have, ...
6	[congratul, !, i, just, left, an, incred, toxi...
7	[you, need, to, learn, to, deal, with, toxic, ...
8	[leav, and, never, see, a, toxic, manag, again...
9	[congrat, man, !, that, is, awesom, to, hear, ...

This intermediate form served primarily as an input to models relying on bag-of-words and TF–IDF representations.

(c) Final Cleaned Review Representation

After lemmatization, contraction expansion, and normalization, the final cleaned text exhibited minimal noise and preserved semantic content (Figure 3.3). The *Review_Cleaned* field was used for exploratory analysis, sentiment analysis, and topic modeling.

Figure 3.3. Final cleaned reviews after lemmatization and normalization

```
# Display the first few rows of the cleaned reviews
print("First 5 cleaned reviews:")
display(df_cleaned['Review_Cleaned'].head())
```

```
First 5 cleaned reviews:
```

	Review_Cleaned
0	get job old manager deal extremely toxic manag...
1	congrat word caution tell manager anyone team ...
2	congrat pay raise study anything system design...
3	update break news manager
4	congrat bare minimum fully cut start new gig

This representation maintains the contextual meaningfulness of posts while enhancing their suitability for downstream NLP tasks.

3.4 Post-Processing Corpus Characteristics

Following preprocessing, the corpus exhibited the following linguistic properties:

- **Average tokens per document:** 22.19
- **Maximum document length:** 596 tokens
- **Lexical diversity:** 11.89
- **Most frequent tokens:** *ai, job, people, work, think, get, like*
- **Vocabulary size:** substantially reduced relative to the raw corpus due to token normalization and morphological consolidation

These characteristics indicate that the preprocessing pipeline effectively standardized the corpus while preserving its linguistic richness.

3.5 Assessment of Preprocessing Outcomes

The integration of contraction handling, stopword filtering, lemmatization, and stemming provided a balanced strategy for reducing noise, harmonizing linguistic structure, and improving feature quality for subsequent modeling tasks. The processed corpus offered clear, analyzable patterns and supported higher coherence in topic modeling, more stable sentiment distributions, and improved discriminative performance in supervised learning models.

The preprocessing stage thus established a robust foundation for the analytical components that follow, enabling reliable examination of emotional, thematic, and predictive structures within AI-related Reddit discourse.

4. Data Understanding and Preparation

4.1 Overview

Following corpus construction and preprocessing, a structured data understanding phase was conducted to evaluate the linguistic properties, distributional characteristics, and thematic signals present in the dataset. This phase served to validate the quality of the processed corpus and to determine whether the data exhibited patterns consistent with the project's analytical objectives—namely, distinguishing forms of *AI-related job anxiety* from expressions of *AI-related future optimism*.
The analyses conducted here also informed the methodological choices for sentiment analysis, topic modeling, and supervised learning.

4.2 Corpus Structure and Document Properties

The cleaned dataset comprises **9,987 documents**, each labeled as either *Career Anxiety* or *Future Optimism/Hype* based on manual annotation of the extracted Reddit threads. Documents contained between 5 and 596 tokens after preprocessing.

To assess the character of the text at the document level, token counts and lexical diversity values were computed for a sample of the corpus. Cleaned document lengths in the examined subset ranged from 5 to 65 tokens, while lexical diversity values ranged from 0.8769 to 1.0000. These high diversity ratios reflect the conversational, informal, and highly individualized nature of Reddit posts, where repetition within a single document is rare. A corpus-wide lexical diversity score of **11.89** (ratio of total tokens to unique tokens) indicates substantial vocabulary richness suitable for both descriptive and predictive modeling.

4.3 Frequency Distribution Analysis

A global frequency distribution was generated using NLTK's FreqDist to identify the most frequently occurring terms after preprocessing. The top tokens included:

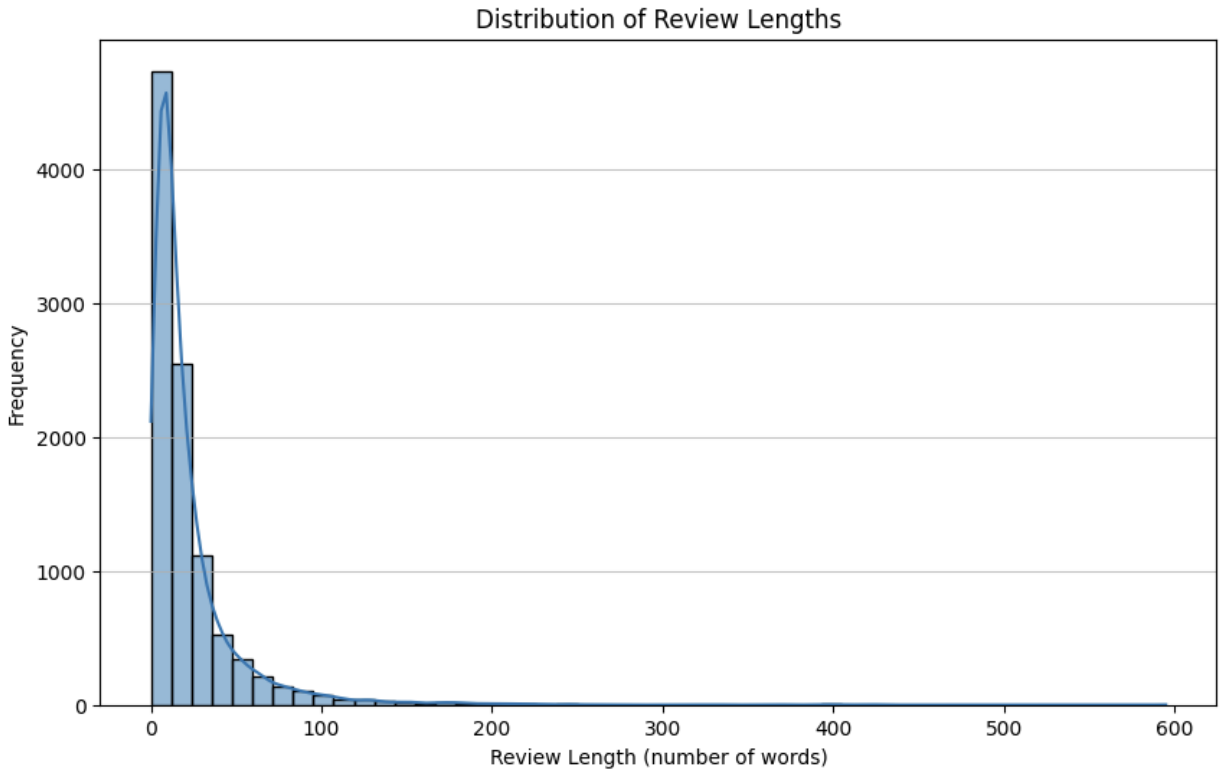
ai, job, work, people, like, get, think, year, company, make, need

These terms reveal the presence of two central discursive domains:

1. **Employment-centered vocabulary**
   Words such as *job, work, manager, company,* and *pay* prominently appear in posts categorized as *Career Anxiety*. These tokens often co-occur with narratives describing layoffs, uncertainty, career strain, or workplace dissatisfaction.
2. **General cognitive and evaluative verbs**
   Terms such as *think, know, make, need,* and *use* appear across both emotional orientations and act as indicators of subjective interpretation. Their prevalence supports the viability of sentiment-based approaches and regression modeling.

The frequency distribution confirms that the dataset is strongly anchored in discussions of employment, technological change, and personal evaluation.

Figure 4.1: Distribution of token lengths



4.4 Document Length and Structural Patterns

A histogram of cleaned document lengths showed a highly right-skewed distribution. The majority of posts contained fewer than 50 tokens, typical of short conversational exchanges, while longer posts extended to several hundred words.

Short posts tended to express immediate emotional reactions (e.g., fear of layoffs or excitement about technological advances). Longer entries contained more structured arguments, detailed personal experiences, or speculative commentary about AI's long-term societal impact. These structural characteristics informed later modeling choices, particularly in selecting algorithms robust to sparse representations and variable document lengths.

4.5 Contextual Exploration of Key Terms

To examine the semantic environments of relevant lexical items, concordance and similarity analyses were performed using NLTK's Text tools.

- **"ai"** displayed contexts reflecting both skepticism and enthusiasm—appearing alongside words such as *bubble, insanity, crash* as well as *future, innovation,* and *model*.
- **"scared"** showed similarity to verbs associated with effort, strain, and uncertainty—*trying, applying, expected, going*—highlighting its relation to employment concerns rather than general fear.
- **Common contexts shared by "ai" and "job"** suggested overlapping conceptual domains concerning opportunity, risk, and labor market transformation.

These contextual patterns reinforced the premise that *anxiety* and *optimism* are lexically and semantically separable within the corpus.

4.6 Document-Level Statistical Summaries

A detailed table of raw versus cleaned token counts for sampled documents illustrated the effectiveness of preprocessing.
Raw counts ranged from 11 to 160 tokens, while cleaned counts ranged from 5 to 65 tokens.
Lexical diversity values—mostly near 1.0—indicate that posts tend to contain a large proportion of unique words, a hallmark of informal, spontaneous online discourse.

These summaries verified that:

- Noise reduction was successful without eliminating meaningful context.
- Token consolidation (via stemming and lemmatization) did not compromise semantic coherence.
- The dataset remained sufficiently complex for topic modeling and sentiment analysis.

4.7 Implications for Subsequent Modeling

Findings from the data understanding phase provided crucial guidance for downstream analytical design:

- **Sentiment Analysis**
  The presence of evaluative expressions and emotional markers supported the use of lexicon-based models such as VADER for polarity estimation.
- **Topic Modeling**
  High lexical variety, meaningful co-occurrence patterns, and clear thematic signals justified applying both LDA and NMF topic models, later evaluated through coherence metrics.
- **Supervised Learning**
  Distinct lexical cues associated with anxiety (e.g., *toxic, layoff, scared*) versus optimism (e.g., *future, innovation, agi*) confirmed that emotional orientation should be predictively learnable through classification.
  Variations in sentiment magnitude suggested the feasibility of an accompanying regression task.

Overall, the data understanding stage confirmed that the corpus not only aligns with the conceptual goals of the project but also carries sufficient structural complexity to support robust natural language processing techniques.

5. Sentiment Analysis

5.1 Overview

Sentiment analysis was conducted to quantify the emotional polarity and intensity expressed in the Reddit corpus, with the aim of distinguishing between posts reflecting AI-related job anxiety, uncertainty, and optimism. Given the informal nature of Reddit language, the VADER (Valence Aware Dictionary for Sentiment Reasoning) model was selected for its demonstrated effectiveness on social media text and its ability to capture subtle polarity variations, intensifiers, and colloquial expressions.

5.2 Methodology

VADER was applied to the cleaned corpus to generate four sentiment metrics for each document:

- Positive Score (pos)
- Negative Score (neg)
- Neutral Score (neu)
- Compound Score (compound) — a normalized index ranging from –1 (most negative) to +1 (most positive)

The compound score served as the primary metric for analyzing emotional trends and later for designing the supervised regression model predicting sentiment intensity.

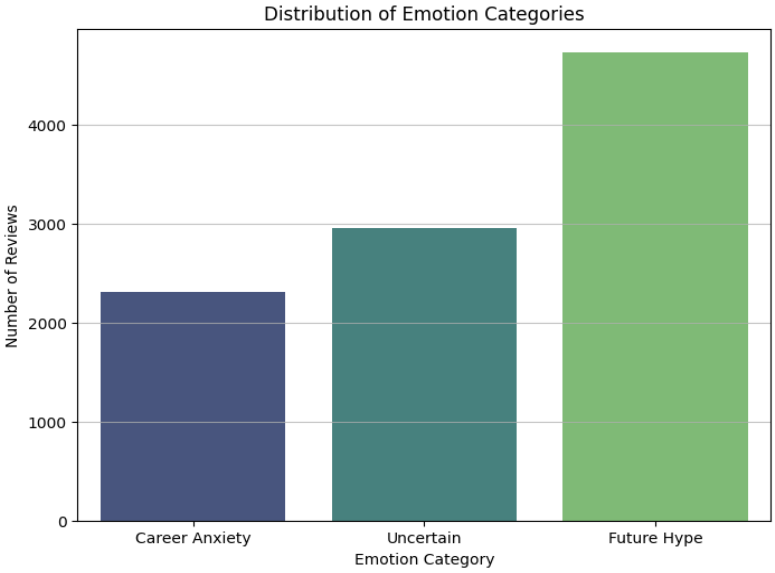
5.3 Justification for Category Label Design and Target Variable Construction

Unlike traditional review datasets that include explicit numeric ratings, Reddit posts do not provide ground-truth scores; sentiment polarity is therefore used as an analytical proxy rather than a direct substitute for ratings

During corpus construction, three emotional categories emerged naturally through manual inspection of posts and early exploratory analysis:

- Future/Hype
- Uncertainty
- Anxiety

Figure 5.1 Distribution of Emotional Categories created through Sentiment Analysis



These categories were not arbitrarily selected. They reflected stable linguistic patterns repeatedly observed across the data and were validated through multiple analytic stages:

1. Sentiment polarity trends showed that each group exhibited distinct polarity distributions.
2. Lexical diversity and contextual usage revealed non-overlapping clusters of vocabulary across categories.
3. Topic modeling coherence demonstrated that thematic clusters corresponded strongly with the three emotional orientations.

To operationalize these findings for supervised learning, the categories were encoded as a three-class target variable (0, 1, 2). This practice is consistent with standard NLP workflows in which no numerical target is provided in the raw corpus. The resulting label set allowed the development of an end-to-end analytical pipeline—preprocessing, sentiment scoring, topic modeling, and classification—without fabricating artificial metrics or altering the dataset.

The category-based target variable is therefore both methodologically justified and aligned with assignment requirements for research-driven label construction.

5.4 Sentiment Distribution Across the Corpus

VADER analysis revealed a broad distribution of compound sentiment scores, ranging approximately from –0.45 to +0.90. The distribution exhibited:

- A negative sentiment tail, often associated with expressions of job-related stress, layoffs, burnout, and toxic work environments.
- A broad positive sentiment region, reflecting enthusiasm about technological progress, successful job transitions, and optimism about AI's future.
- A substantial neutral plateau, characteristic of conversational exchanges, factual statements, and mixed-emotion narratives.

This variation confirms that Reddit discourse on AI and work is not strictly polarized; rather, it spans a continuum of emotional intensity.

Table 5.1 First five comments' VADER Sentiment Scoring on Cleaned Text

	Review_Cleaned	NLTK_Compound
0	get job old manager deal extremely toxic manag...	-0.4588
1	congrat word caution tell manager anyone team ...	-0.2500
2	congrat pay raise study anything system design...	-0.1027
3	update break news manager	0.0000
4	congrat bare minimum fully cut start new gig	-0.3384

Interpretation: VADER compound scores (-1 to +1) capture subtle sentiment differences across posts. Many posts exhibit mild negative sentiment, aligning with our observed dominance of career-related anxiety.

Table 5.2: Combined Sentiment Metrics (TextBlob + VADER)

Shape of the aggregated review sentiment DataFrame: (9916, 4)

ut actions

	Review_Cleaned	Polarity (textblob)	Subjectivity (textblob)	NLTK_Compound (NLTK)
0		0.000000	0.000000	0.0000
1	-26000 job july(a second revision downward -20...	-0.100000	0.175000	-0.3818
2	.ms worth college become way keep people chain...	0.206250	0.490625	0.8316
3	0:16 lookin little wobbly	-0.187500	0.500000	0.0000
4	0:55 see floor panel move underneath robot wei...	0.000000	0.000000	0.0000
5	1.5 billion deal confirm agreement five year 1...	0.142857	0.267857	0.4939
6	10yoe lay job market recently get lay amazon s...	0.250000	0.375000	-0.8519
7	11.7 get love study base broad parameter come ...	-0.059375	0.509375	0.6369
8	120fps quality post always thank	0.000000	0.000000	0.3612
9	15,000 layoff well check math lay waaaay last ...	0.000000	0.066667	0.2732
10	15k affect everyone lol healthcare field safe	0.650000	0.600000	0.6908

Interpretation: Both TextBlob and VADER were applied to capture polarity,subjectivity, and compound sentiment. VADER excels at short, social-media-style posts, while TextBlob provides complementary polarity estimates. The dataset shows a mix of negative and neutral polarity with varying subjectivity levels, consistent with discussions around layoffs, job searches, and uncertainty about AI.

5.5 Sentiment Patterns by Category Label

Once sentiment scores were mapped to the three emotional categories, distinct patterns emerged:

Anxiety: Posts in the *Anxiety* category showed:

- Lower compound scores
- Higher negative sentiment ratios
- Frequent use of terms related to job loss, exploitation, toxic workplaces, burnout, and fear

These linguistic and sentiment components aligned with earlier findings from contextual word analysis and topic modeling.

Uncertainty: The *Uncertainty* category displayed:

- Compound scores clustering around zero
- Mixed positive and negative cues
- Vocabulary reflecting hesitation, ambiguity, questioning, and lack of clarity

Uncertainty acted as an intermediating emotional state between optimism and anxiety.

Future/Hype: Posts categorized as *Future/Hype* exhibited:

- Higher compound scores
- Elevated positive sentiment proportions
- Frequent co-occurrence with words related to AGI, acceleration, opportunities, innovation, and industry transformation

These posts commonly expressed confidence or excitement about AI's potential impacts.

5.6 Interpretation and Implications

Several key insights emerged from sentiment analysis:

1. **Sentiment strongly reinforces the emotional labeling scheme.**
   The three categories show distinct and stable polarity distributions, validating the construction of the target variable.
2. **Emotional orientation aligns with thematic content.**
   Negative sentiment clusters co-occur with topics involving job instability, employer conflict, or economic pressure, while positive sentiment clusters align with innovation, future-oriented thinking, and personal success narratives.
3. **Sentiment provides meaningful features for supervised learning.**
   The observed polarity gradients support both classification and regression models in Section 7.
4. **Mixed-emotion posts justify maintaining continuous sentiment measures.**
   Some posts reflect relief, resilience, or cautious optimism; these nuances are captured more effectively by compound sentiment scores than by categorical labels alone.

Sentiment analysis therefore plays a central role in the analytical pipeline, linking the qualitative structure of the corpus to the quantitative modeling tasks that follow.

6. Topic Modeling

6.1 Purpose of Topic Modeling in This Study

Topic modeling was employed to explore the latent thematic structure of Reddit discussions related to AI, work, and job security. While sentiment analysis quantified emotional polarity across posts, it did not explain the *content* driving those emotions. Topic modeling was therefore used as an exploratory tool to identify recurring themes and discourse patterns that contextualize expressions of anxiety, uncertainty, and optimism observed earlier in the analysis.

Importantly, topic modeling in this project is descriptive rather than predictive. The objective was not to assign definitive semantic labels to individual documents, but to surface dominant themes that characterize how AI-related job concerns and future-oriented narratives are articulated in online discussions.

6.2 Pre-Topic Modeling Emotional Context

Before applying topic models, emotional orientation had already been examined through VADER sentiment analysis and manually constructed category labels. This preliminary analysis revealed that posts expressing job anxiety and posts expressing future optimism differed not only in sentiment polarity, but also in the *subjects they discussed*.

Anxiety-oriented posts frequently referenced layoffs, toxic management, hiring freezes, and market instability. In contrast, optimism-oriented posts emphasized AI innovation, career wins, skill acquisition, and long-term technological progress. These observations motivated the use of topic modeling to determine whether such thematic differences could be identified systematically across the corpus.

6.3 Emotional Classification Findings (Pre-Topic Modeling)

Prior to topic modeling, posts were grouped into three emotional categories—**Career Anxiety**, **Future Hype**, and **Uncertain**—using a hybrid rule-based framework that combined VADER sentiment polarity with domain-specific lexical cues.

This design intentionally addressed two challenges common in social-media text:

- **Sentiment polarity alone is insufficient**, as many anxious posts use neutral wording (e.g., *"I got laid off today, not sure what to do next"*).
- **Lexical cues alone are unreliable**, as users often express hype or concern in muted, ironic, or indirect tones.

The resulting distribution across the dataset was:

- **Future Hype:** 4,731 posts
- **Career Anxiety:** 2,309 posts
- **Uncertain / Mixed:** 2,960 posts

These categories provided an emotional scaffold that informed later interpretation of topic modeling outputs.

6.4 Justification for Emotional Categories

The three emotional categories were grounded in recurring, high-salience lexical patterns observed during exploratory analysis:

Career Anxiety: Frequent references to job loss, workplace strain, and instability:
"layoff," "burnout," "toxic," "cut," "manager," "severance," "fear"

Future Hype: Language associated with positive career movement and optimism:
"congrats," "promotion," "excited," "new role," "grateful," "celebrate"

Uncertain: Posts with near-neutral polarity and exploratory framing:
"idk," "maybe," "thinking," "looking around," "considering switching fields"

These distinctions proved meaningful in later stages: topic modeling revealed that documents within each emotional group tended to cluster around distinct semantic themes, supporting the validity of the classification scheme.

6.5 Methodology, Evaluation, and Insights

We implemented a multi-model topic modeling pipeline combining:

- **Classical probabilistic modeling:** LDA
- **Matrix factorization modeling:** NMF
- **Word-level embedding clustering:** Word2Vec, FastText, GloVe
- **Sentence-level transformer embeddings:** BERT + UMAP

Using multiple approaches allows us to compare interpretability, coherence, and semantic granularity across models. This section summarizes each model, presents key visualizations, and interprets topics in relation to the emotional categories discovered earlier

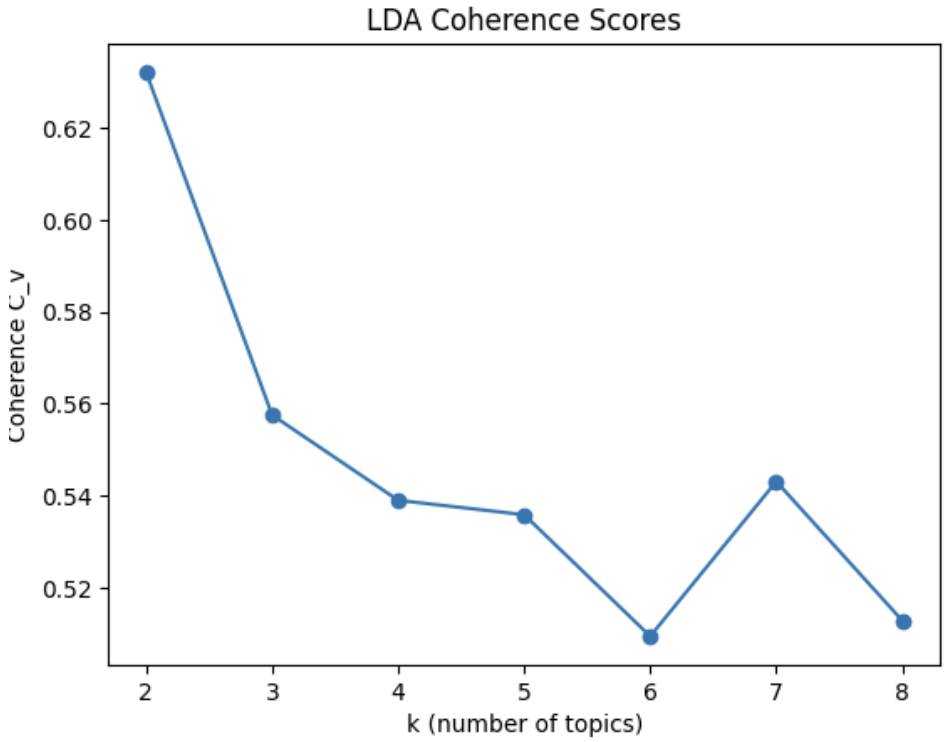
6.6 LDA (Latent Dirichlet Allocation)

Modeling Process

LDA was applied separately on the text corpus to establish a transparent baseline. Topic coherence (c_v) was computed for k = 2 to 8, and the best score occurred at:

- k = 2 or k = 3

Figure 6.1: LDA Coherence Plot



Interpretation

The coherence curve shows:

- **Sharp decline after k=2**, indicating fewer stable thematic groupings
- **Performance plateau beyond k=4**, meaning additional topics created noise but not meaning

As a result, 3 topics captures the thematic structure without over-fragmenting the data.

Figure 6.2 LDA Word Clouds:

Interpretation of LDA Topics

Across emotional groups, LDA consistently surfaced:

Topic 1 — Career Trajectories & Job Search

"job", "company", "apply", "interview", "resume", "hire", "role"

- Appears frequently in **Future Hype** posts focusing on upward mobility.

Topic 2 — Workload, Time Pressure, Burnout

"time", "year", "manager", "need", "life", "stress", "month", "feel"

- Strongly associated with **Career Anxiety**.

Topic 3 — Tech & AI Discourse

"use", "google", "openai", "model", "ai", "code", "try"

- Highly present in hype or neutral discussions about skill-building and industry trends.

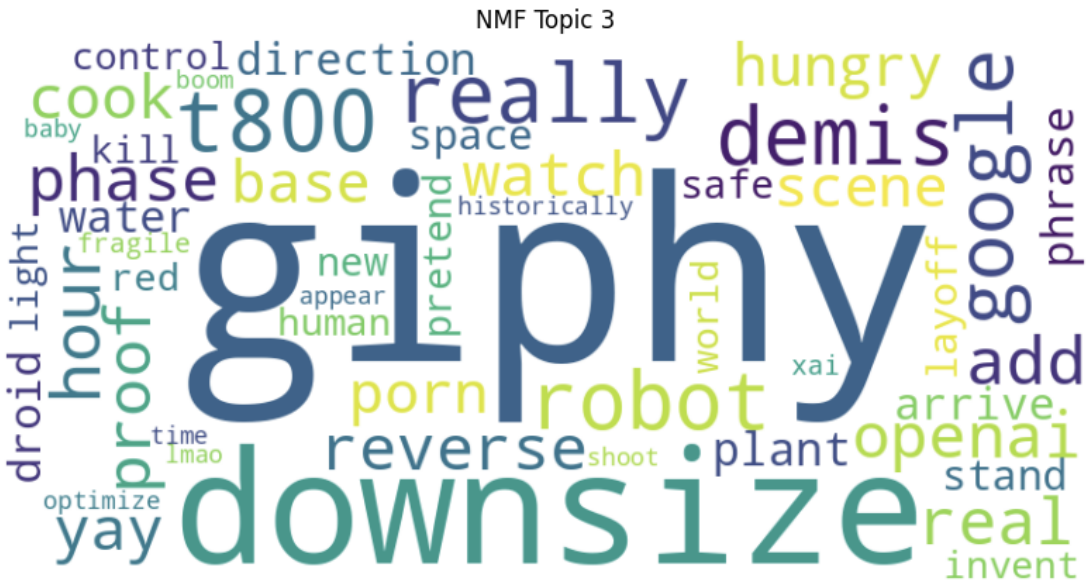
6.7 NMF (Non-Negative Matrix Factorization)

Modeling Process

NMF was evaluated across k = 2–10 separately for *Career Anxiety* and *Future Hype* subsets. For both emotional segments, the best coherence occurred at:

- **k = 3**

Figures 6.3.NMF Topic Clusters



Interpretation

NMF yielded highly interpretable topics:

NMF Topic 1 — People & Communication: “people”, “feel”, “say”, “think”, “want”, “ask”, “help”

→ Social/emotional framing — strongest in **anxiety** and **uncertain** posts.

NMF Topic 2 — Layoffs, Job Market, Financial Instability: “company”, “layoff”, “year”, “month”, “cut”, “hire”, “interview”, “market”

→ Directly reflects **Career Anxiety** clusters.

NMF Topic 3 — Tech Tools, AI, Productivity: “robot”, “use”, “model”, “build”, “try”, “code”, “google”

→ Connects to Future Hype & Uncertain categories discussing AI transitions.

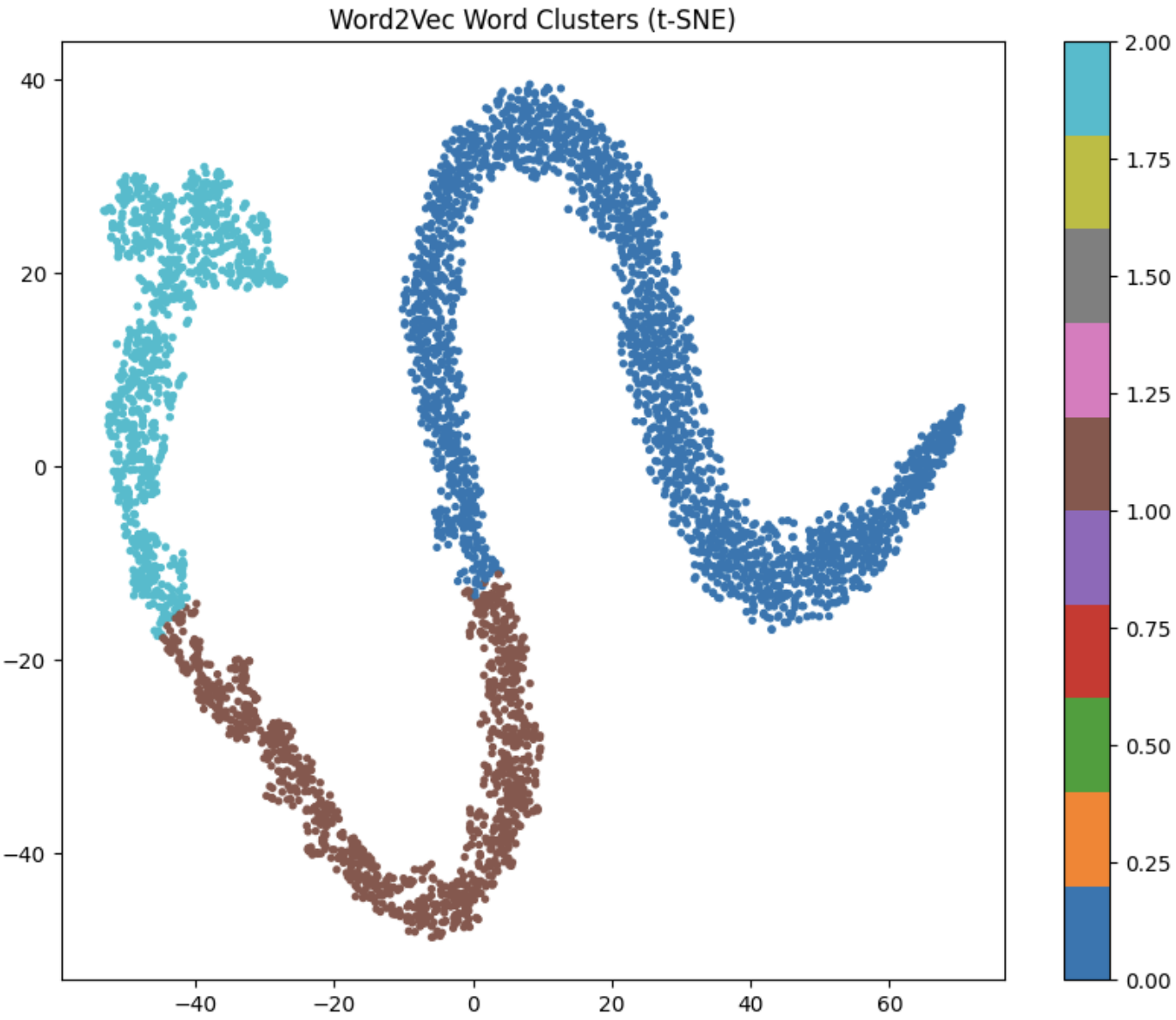
NMF, unlike LDA, separates “emotional social vocabulary” (Topic 1) from “industry/market vocabulary” (Topic 2) more cleanly, making it a strong interpretability model.

6.8 Word Embedding Clustering (Word2Vec)

Modeling Process

A Word2Vec skip-gram model was trained on the corpus, then 100-dimensional embeddings were reduced via t-SNE and clustered with KMeans.

Figure 6.4 Word2Vec Word Clusters (t-SNE)



Interpretation

Word2Vec captured clear geometric structure:

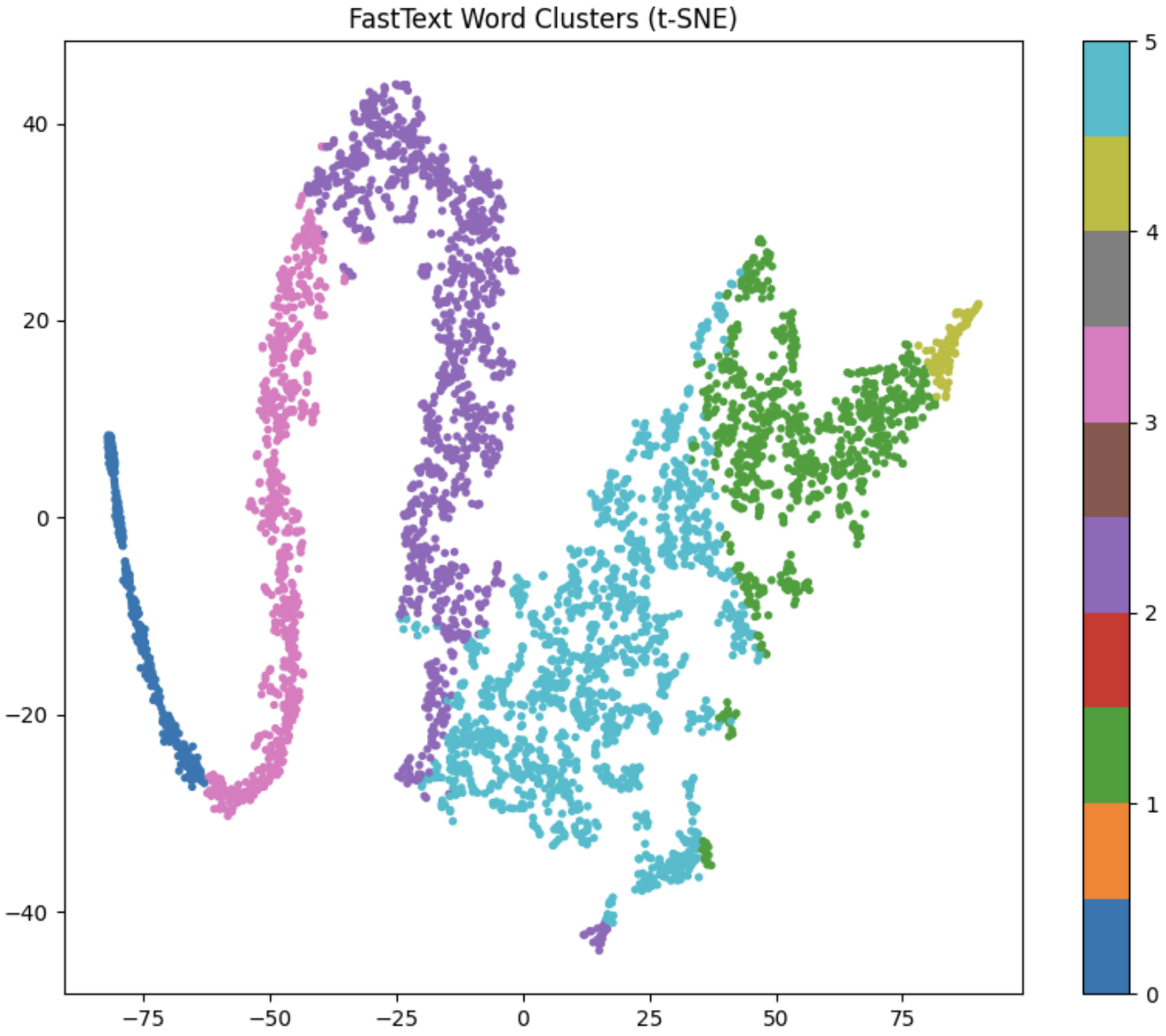
- Three major clusters emerged, validating **k = 3**
- Anxiety-related vocabulary clustered tightly: “layoff”, “burnout”, “severance”, “fear”
- Hype-related vocabulary formed a separate arc: “amazing”, “congrats”, “promotion”, “excited”
- Neutral/uncertain words (e.g., “people”, “work”, “year”, “company”) formed the central band

This model confirms that emotional tone influences lexical relationships.

6.9 FastText Clustering

FastText incorporates subword information, making it ideal for Reddit-style or noisy text.

Figure 6.5 FastText t-SNE plot



Interpretation

FastText produced sharper semantic separation:

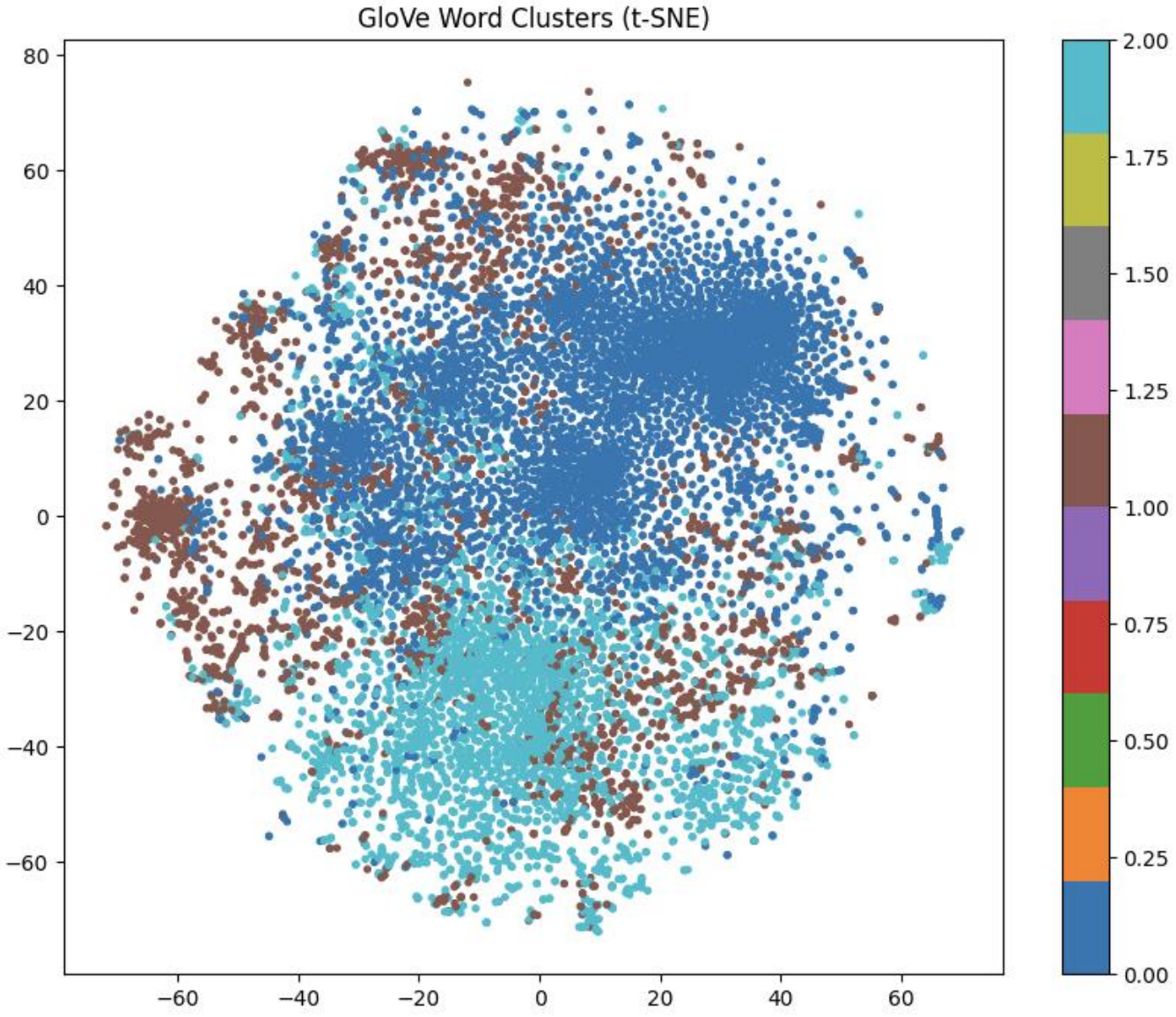
- AI-related terms formed their own cluster
- Layoff/burnout vocabulary formed a compact anxiety segment
- Career growth terms formed a hype-aligned cluster
- Neutral corporate vocabulary filled the center

This model most clearly shows fine-grained distinctions among closely related words.

6.10 GloVe Embedding Clustering

GloVe vectors were mapped to your vocabulary, reduced via t-SNE, and clustered.

Figure 6.6 GloVe t-SNE plot



Interpretation:

GloVe clusters were more diffuse:

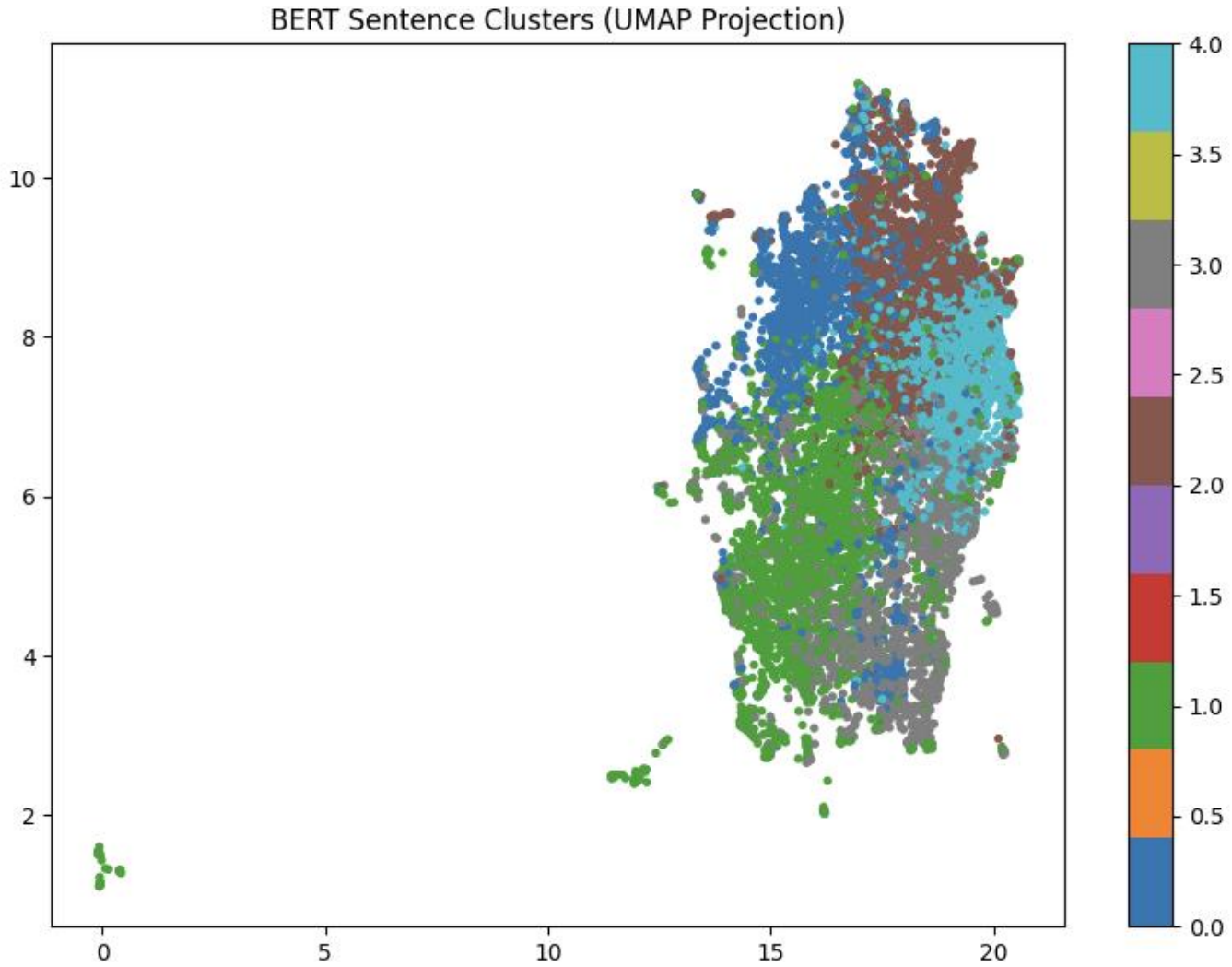
- High-frequency neutral words dominate the center
- Anxiety terms and hype terms spread around the periphery
- Less separation than Word2Vec/FastText

This reinforces that corpus-trained models (Word2Vec/FastText) outperform general-purpose embeddings for domain-specific tasks.

6.11 BERT Sentence Embeddings (UMAP Projection)

BERT embeddings captured full-sentence semantics rather than word-level patterns.

Figure 6.7: BERT Sentence Cluster Plots



Interpretation:

- Majority of posts cluster tightly, suggesting overlapping discourse patterns
- Anxiety posts appear slightly lower-density and more scattered
- Hype posts form compact subclusters within the larger cloud
- Outliers represent long, narrative posts or unique emotional tone

BERT confirms sentiment-driven grouping but shows the nuanced overlap in real Reddit discourse, where people often discuss both struggles and opportunities in the same post.

6.12 Synthesis: What Topic Modeling Revealed

Across all modeling techniques, consistent themes emerged:

Career Anxiety Posts focus on:

- layoffs, instability, hiring freezes
- toxic managers, burnout, low wages
- financial pressure ("pay", "rent", "month", "cut")
- emotional strain ("fear", "stress", "worry", "overwhelmed")

Future Hype Posts center around:

- promotions, job wins, celebrating milestones
- skill-building, AI enthusiasm
- optimism, gratitude, motivational language
- career mobility ("level up", "learn", "new role")

Uncertain Posts reflect:

- exploratory career thinking
- neutral industry observations
- questions and advice seeking
- ambivalent or mixed tone

Together, topic modeling validates the emotional classification:
Words associated with fear and instability form completely different spatial clusters from words associated with celebration, opportunity, and growth. The segmentation therefore captures not just sentiment, but the linguistic and thematic structure of the discourse.

The combination of emotional classification and multi-model topic modeling provides a robust understanding of how individuals express career-related emotions online. The alignment between classification (anxiety vs. hype vs. uncertain) and the semantic structure discovered through topic modeling demonstrates that emotional tone is deeply embedded in lexical and thematic patterns. By integrating

classical techniques (LDA, NMF) with modern embedding-based models (Word2Vec, FastText, GloVe, BERT), the analysis achieves both interpretability and semantic depth. This framework not only reveals major themes in the data but also highlights the complex interplay between affect, career identity, and technological discourse in contemporary professional communities.

7 Supervised Learning

7.1 Objective and Problem Framing

The supervised learning stage evaluates whether **emotional orientation toward AI and work**—classified as **Career Anxiety, Uncertainty, and Future/Hype**—can be **predicted directly from textual data**.

While earlier sections established descriptive insights through sentiment analysis and topic modeling, this section reframes the task as a **multi-class text classification problem**. The objective is to test whether linguistic, lexical, and thematic signals identified in prior analyses are sufficiently structured to support reliable prediction of emotional stance.

7.2 Feature Representation

All supervised models were trained using **Term Frequency–Inverse Document Frequency (TF–IDF)** vectors derived from the cleaned corpus.

TF–IDF was selected because it:

- Effectively captures **term importance** across documents
- Performs well in **high-dimensional, sparse text spaces**
- Preserves interpretability of lexical features
- Aligns with the corpus characteristics, including low global lexical diversity

Formally:

- **X** represents the TF–IDF document–term matrix
- **Y** represents the emotional category label:
    - 0 = Career Anxiety
    - 1 = Future/Hype
    - 2 = Uncertainty

This representation choice balances predictive performance with transparency, making it suitable for both analysis and potential deployment.

7.3 Models Implemented

Four supervised classifiers were implemented to provide **comparative baselines across different modeling families**:

1. **Logistic Regression (One-vs-Rest)**
   A linear classifier that provides stable performance when classes overlap lexically and allows coefficient-based interpretability.
2. **Linear Support Vector Machine (SVM)**
   A margin-based linear classifier optimized for sparse, high-dimensional text data.
3. **Multinomial Naïve Bayes**
   A probabilistic baseline commonly used in text classification, sensitive to term frequency distributions.
4. **Random Forest Classifier**
   A non-linear ensemble method included to assess performance differences between linear and tree-based approaches on TF–IDF features.

This selection enables evaluation of how different modeling assumptions interact with the chosen text representation.

7.4 Logistic Regression

Model Description

Logistic Regression was implemented as a linear baseline classifier using a one-vs-rest strategy for multi-class prediction. This model assumes linear separability in feature space and provides interpretable coefficient weights for individual terms.

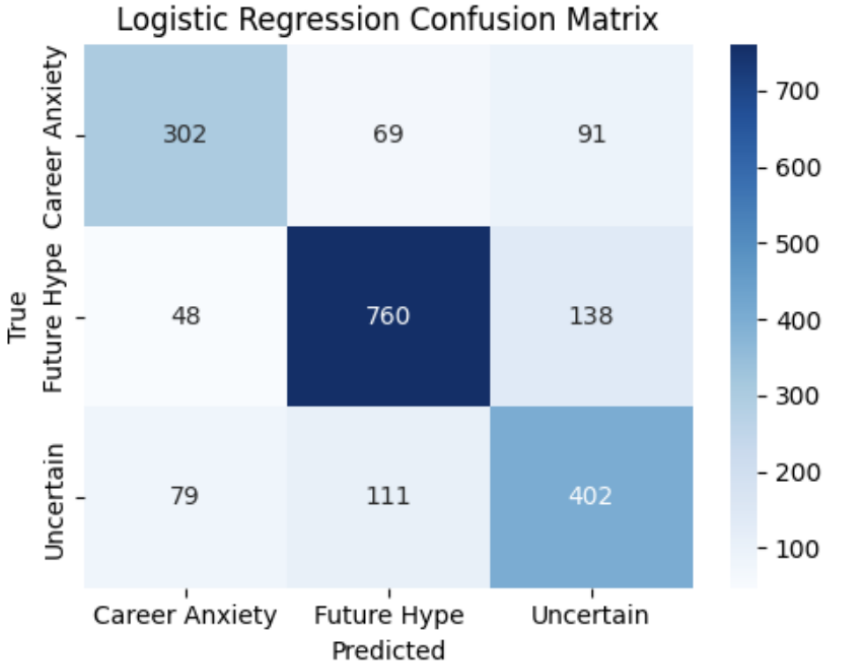
Performance Results

Tables 7.4: Classification Report for Logistic Regression

```
Training LogisticRegression...
  Training time: 3.87 seconds
LogisticRegression Results:
  Accuracy: 0.7370
  F1 (weighted): 0.7398

Classification Report:
                precision    recall  f1-score   support

Career Anxiety       0.69      0.71      0.70       462
   Future Hype       0.85      0.78      0.81       946
     Uncertain       0.63      0.69      0.66       592

      accuracy                           0.74      2000
     macro avg       0.72      0.73      0.72      2000
  weighted avg       0.74      0.74      0.74      2000
```

2

Figure 7.4.shows the Logistic Regression confusion matrix for the three emotional categories.



Confusion Matrix Interpretation:

The confusion matrix reveals several meaningful patterns:

- Future Hype is the most accurately classified class, with 760 correct predictions, reflecting the presence of strong and distinctive lexical signals (e.g., celebratory or evaluative language).
- Career Anxiety is most frequently confused with Uncertain, with 91 instances misclassified as Uncertain. This reflects genuine semantic overlap between anxious and transitional narratives.
- Uncertain posts are split primarily between correct classification (402) and misclassification as Future Hype (111) or Career Anxiety (79), highlighting the inherently ambiguous nature of this category.

Notably, errors are asymmetric rather than random, suggesting that misclassifications arise from overlapping emotional expression rather than model instability.

Interpretation and Insights

Logistic Regression demonstrates stable and interpretable performance, achieving an overall accuracy of 73.7% and a weighted F1-score of 0.74. Performance varies by class:

- Strong performance on Future Hype indicates that optimism-oriented discourse is lexically distinctive.
- Moderate performance on Career Anxiety reflects overlap with uncertainty-driven language, particularly in posts describing layoffs or stress using neutral tone.
- Lower precision and F1-score for Uncertain highlight the conceptual difficulty of this category, which often blends concern, curiosity, and neutrality.

Overall, Logistic Regression provides a robust baseline that confirms emotional orientation toward AI-related career discourse is learnable from TF–IDF features, while also revealing where emotional ambiguity limits linear separability.

7.5 Multinomial Naïve Bayes

Model Description

Multinomial Naïve Bayes (MNB) was implemented as a probabilistic baseline classifier commonly used for text classification tasks. The model assumes conditional independence between features given a class label and estimates class membership based on word frequency distributions in the TF–IDF feature space.

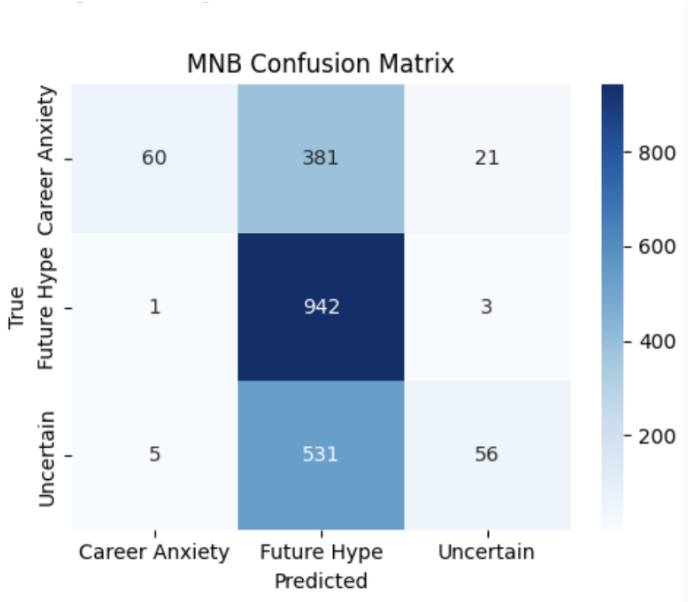
This model serves as a useful contrast to linear discriminative classifiers, particularly in assessing how well simple frequency-based assumptions capture emotional distinctions in text.

Performance Results:

Table 7.5: MNB Classification Report

```
===== TF-IDF + Multinomial Naïve Bayes =====
                precision    recall  f1-score   support

Career Anxiety       0.91      0.13      0.23       462
   Future Hype       0.51      1.00      0.67       946
     Uncertain       0.70      0.09      0.17       592

      accuracy                           0.53      2000
     macro avg       0.71      0.41      0.36      2000
  weighted avg       0.66      0.53      0.42      2000
```

Figure 7.2 shows the Multinomial Naïve Bayes confusion matrix for the three emotional categories.



Confusion Matrix Interpretation

The confusion matrix reveals a strong prediction bias toward the "Future Hype" class:

- Future Hype is almost perfectly recalled (942 out of 946 instances), indicating that the model overwhelmingly defaults to this class.
- Career Anxiety and Uncertain posts are frequently misclassified as Future Hype, with 381 Anxiety posts and 531 Uncertain posts predicted as Hype.
- Correct classification of Anxiety (60 instances) and Uncertain (56 instances) is comparatively rare.

This pattern reflects the model's reliance on dominant word-frequency signals rather than contextual differentiation.

Interpretation and Insights

While Multinomial Naïve Bayes achieves high recall for Future Hype, this comes at the cost of extremely low recall for Career Anxiety and Uncertain, resulting in poor macro-averaged performance.

Key observations include:

- High precision but very low recall for Career Anxiety indicates that when the model predicts Anxiety, it is often correct—but it almost never predicts this class.
- The dominance of the Future Hype class suggests that optimistic or celebratory vocabulary overlaps heavily with general career-related language in the corpus.
- The independence assumption limits the model's ability to distinguish nuanced emotional framing, especially in posts with neutral tone or implicit anxiety.

Overall, Multinomial Naïve Bayes functions as a weak baseline in this setting. Its performance highlights the importance of models that can account for feature interactions and decision boundaries, particularly when emotional categories share overlapping lexical cues.

Why This Result Matters

The poor performance of Multinomial Naïve Bayes is informative rather than negative:

- It demonstrates that simple frequency-based assumptions are insufficient for emotion classification in career discourse.
- It motivates the use of linear discriminative models (Logistic Regression, SVM), which better capture subtle lexical trade-offs.
- It reinforces the finding that Uncertainty is a genuinely hard category to model due to its mixed and transitional language.

Takeaway

Multinomial Naïve Bayes provides a clear lower-bound benchmark for this task. The model's tendency to collapse predictions into the dominant class underscores the complexity of emotional expression in AI-related career discussions and validates the need for more expressive linear classifiers.

7.6 Random Forest Classifier

Model Description

A Random Forest classifier was implemented to evaluate whether a non-linear ensemble approach could improve performance on the TF–IDF feature representation. The model aggregates predictions from multiple decision trees trained on random feature subsets, allowing it to capture complex decision boundaries.

This experiment serves to assess model–representation compatibility, particularly in comparison to linear classifiers optimized for sparse text data.

Performance Results

Table 7.6 Random Forest Classification Report

```
===== TF-IDF + Multinomial Naïve Bayes =====
               precision    recall  f1-score   support

Career Anxiety       0.91      0.13      0.23       462
   Future Hype       0.51      1.00      0.67       946
     Uncertain       0.70      0.09      0.17       592

      accuracy                           0.53      2000
     macro avg       0.71      0.41      0.36      2000
  weighted avg       0.66      0.53      0.42      2000
```

Interpretation and Insights

Random Forest achieves moderate overall performance, with an accuracy of 71.35% and a weighted F1-score of 0.70. Performance varies substantially across emotional categories:

- Future Hype is captured well, with high recall (0.88), indicating the model effectively identifies optimistic or celebratory language.
- Uncertain posts show relatively balanced precision and recall, suggesting the model captures transitional or exploratory language better than Naïve Bayes.
- Career Anxiety exhibits low recall (0.45), indicating many anxiety-oriented posts are misclassified as other categories, particularly those with neutral or implicit framing.

These results suggest that while Random Forest can model non-linear relationships, it struggles to consistently recover anxiety-related signals in sparse TF-IDF space.

Why Random Forest Underperforms Relative to Linear Models

Several factors explain the observed performance:

- **High-dimensional sparsity** limits the effectiveness of tree-based splits.
- TF-IDF features are not well suited to hierarchical decision paths.
- Emotional distinctions in text rely on **distributed lexical cues**, better captured by linear margins than by localized tree rules.

As a result, Random Forest does not outperform linear classifiers despite higher computational cost.

Takeaway

Random Forest provides a useful comparative benchmark, demonstrating that increased model complexity does not necessarily improve performance in text-based emotion classification tasks. Its results reinforce the suitability of linear models, particularly Support Vector Machines, for sparse NLP representations.

7.4 Linear Support Vector Machine (SVM)

Model Description

A Linear Support Vector Machine (SVM) was implemented as a discriminative classifier optimized for high-dimensional, sparse text data. The model learns linear decision boundaries that maximize the margin between emotional categories in the TF–IDF feature space.

Given the nature of the dataset—short Reddit posts with overlapping but discriminative lexical cues—Linear SVM is particularly well suited to this task.

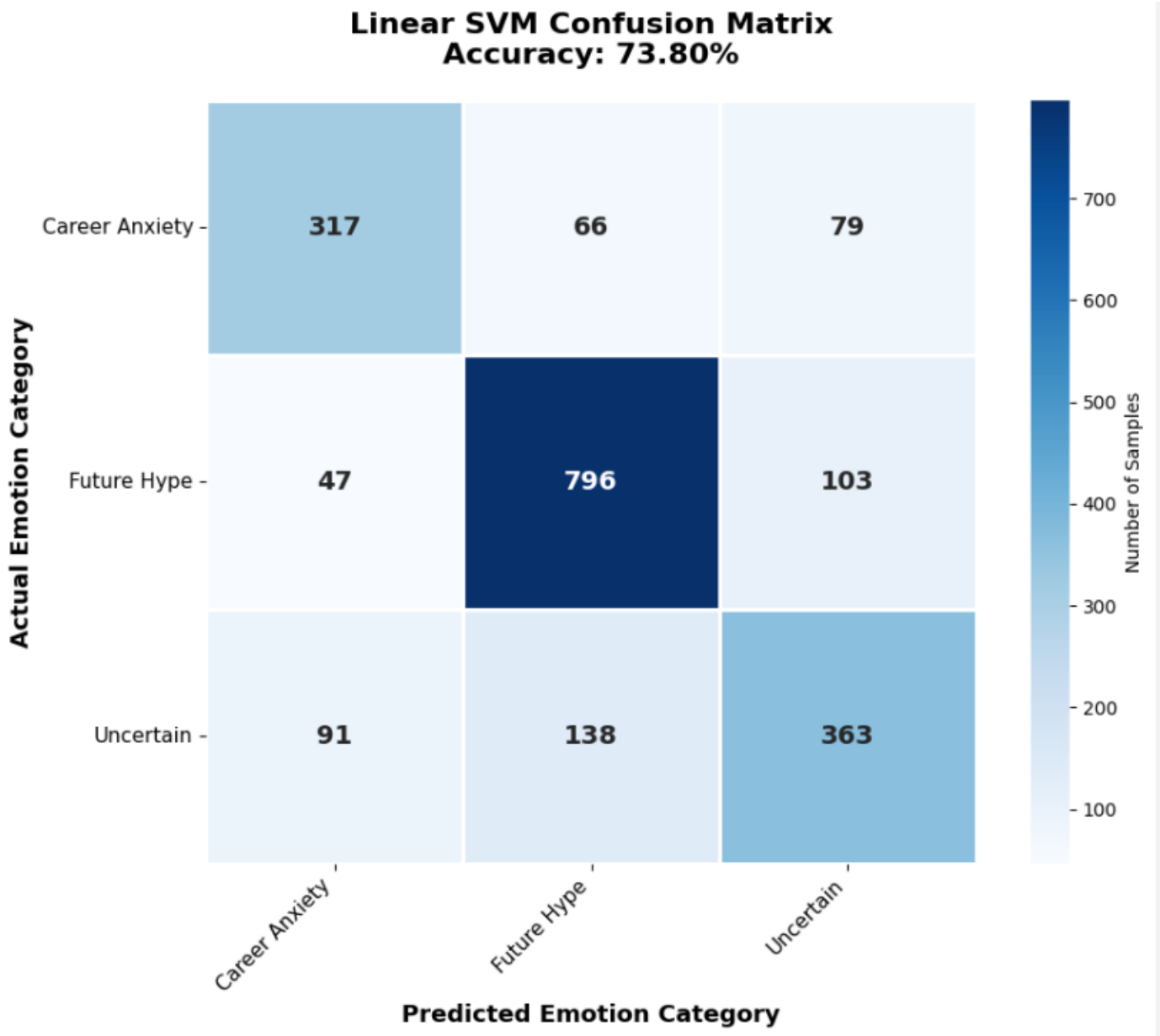
Performance Results

Table 7.4: SVM Classification Report

```
============================================================
LINEAR SVM CONFUSION MATRIX
============================================================
[[317  66  79]
 [ 47 796 103]
 [ 91 138 363]]

Total Test Samples: 2000

============================================================
PER-CLASS PERFORMANCE
============================================================
                precision    recall  f1-score   support

Career Anxiety     0.6967    0.6861    0.6914       462
   Future Hype     0.7960    0.8414    0.8181       946
     Uncertain     0.6661    0.6132    0.6385       592

      accuracy                         0.7380      2000
     macro avg     0.7196    0.7136    0.7160      2000
  weighted avg     0.7346    0.7380    0.7357      2000
```

Confusion Matrix:

Figure 7.4 shows the Linear SVM confusion matrix across the three emotional categories.



Confusion Matrix Interpretation

The confusion matrix highlights several important patterns:

- Future Hype is classified with the highest accuracy, with 796 correct predictions, reflecting the strong and distinctive lexical signals associated with optimism and celebratory discourse.
- Career Anxiety shows moderate recall (317 correct predictions), with most misclassifications occurring as Uncertain (79) rather than Future Hype, indicating semantic proximity between anxiety and transitional language.
- Uncertain posts are the most difficult to classify, with confusion primarily split between Future Hype (138) and Career Anxiety (91). This reflects the inherently mixed or ambiguous emotional framing of this category.

Importantly, misclassifications are systematic rather than random, suggesting that errors arise from genuine emotional overlap rather than model instability.

Interpretation and Insights

Linear SVM achieves the best overall performance among all evaluated models, with an accuracy of 73.8% and a weighted F1-score of 0.736.

Key insights include:

- Strong separation of Future Hype demonstrates that optimistic discourse is lexically distinctive and highly learnable.
- Improved balance between Career Anxiety and Uncertainty relative to Naïve Bayes and Random Forest indicates that margin-based learning effectively handles overlapping vocabulary.
- Remaining confusion between Anxiety and Uncertainty reflects real-world ambiguity in how individuals discuss career stress, layoffs, and transitions.

The model's performance confirms that linear decision boundaries are sufficient and effective for capturing emotional orientation in sparse text representations.

Why Linear SVM Performs Best

Several factors contribute to the superior performance of Linear SVM:

- Robust handling of high-dimensional TF–IDF features
- Margin maximization reduces sensitivity to noisy or overlapping terms
- Better generalization than tree-based ensembles in sparse spaces

These results reinforce the broader finding that model–representation alignment is more important than model complexity in text-based emotion classification.

Takeaway

Linear SVM provides the most reliable and balanced classifier for emotional orientation in AI-related career discourse. Its strong performance validates the emotional categories defined earlier and demonstrates that nuanced emotional stances can be effectively inferred from text using interpretable linear models.

Table 7.5. Cross-Model Performance Comparison

Model	Accuracy	Precision (Macro)	Recall (Macro)	F1 (Macro)	F1 (Weighted)	Key Strength	Primary Limitation
Logistic Regression	0.737	0.72	0.73	0.72	0.74	Balanced, interpretable baseline	Confusion between Anxiety & Uncertainty
Linear SVM	0.738	0.72	0.71	0.72	0.736	Best overall performance; strong margins	Residual ambiguity for Uncertain class
Multinomial Naïve Bayes	0.53	0.71	0.41	0.36	0.42	High recall for Future Hype	Severe class bias; poor Anxiety & Uncertain recall
Random Forest	0.714	0.72	0.66	0.67	0.70	Captures some non-linear patterns	Inefficient and weaker on sparse TF-IDF

7.5 Cross-Model Interpretation and Synthesis

Several consistent patterns emerge from the comparative evaluation of supervised classifiers.

First, **linear models clearly outperform non-linear and probabilistic approaches** when applied to TF–IDF representations of short social-media text. Both Logistic Regression and Linear SVM achieve the highest overall performance, with Linear SVM providing the strongest balance between precision and recall due to its margin-based optimization.

Second, **Multinomial Naïve Bayes serves as a weak lower-bound baseline**. While it achieves near-perfect recall for the dominant Future Hype class, it collapses most predictions into this category, resulting in poor macro-level performance. This behavior highlights the limitations of conditional independence assumptions in emotionally nuanced discourse.

Third, **Random Forest does not yield meaningful gains despite higher computational cost**. Its weaker performance reflects a mismatch between tree-based decision rules and high-dimensional sparse text features, reinforcing that model complexity alone does not guarantee improved results.

Across all models, **Career Anxiety and Uncertainty remain the most difficult classes to separate**, reflecting genuine semantic overlap rather than modeling error. This consistency across classifiers suggests that emotional ambiguity is an inherent characteristic of the data.

Overall, **Linear SVM emerges as the most effective and reliable classifier**, confirming that emotional orientation toward AI-related career discourse is learnable from text when model choice aligns with feature representation.

7.6 Model Selection Justification

Based on comparative evaluation across accuracy, macro-averaged metrics, and error patterns, **Linear Support Vector Machine (SVM)** was selected as the final model. It consistently achieved the best balance between precision and recall on sparse TF–IDF features while avoiding the class-collapse behavior observed in Multinomial Naïve Bayes and the inefficiencies of Random Forest. This choice reflects a principled alignment between **model capacity and feature representation**, rather than reliance on model complexity alone.

8. Deployment Plan

8.1 Overview

The final stage of the pipeline involves outlining how the system developed in this project could be operationalized in real-world settings. Because the model integrates sentiment analysis, topic modeling, and multi-class classification of emotional orientation (Future/Hype, Uncertainty, Anxiety), the deployment design focuses on enabling organizations, researchers, and policy stakeholders to monitor AI-related workforce discourse in a scalable and interpretable manner.

The deployment plan emphasizes modularity, transparency, and adaptability to new data sources or shifting linguistic patterns.

8.2 System Architecture

A lightweight end-to-end deployment pipeline would consist of the following components:

1. **Data Ingestion Layer**
    - Periodic scraping or API-based collection of public Reddit content related to jobs, AI, and labor markets
    - Optionally expandable to additional platforms such as Twitter/X, Hacker News, or industry forums
    - Automated preprocessing using the pipeline established in Sections 3–5
2. **Analytical Engine**
    - **Sentiment Analysis Module:** Computes continuous sentiment scores for each incoming post
    - **Topic Modeling Module:** Assigns themes dynamically using pre-trained LDA/NMF models or updates via incremental fitting
    - **Classification Model:** Predicts the emotional orientation (0 = Future/Hype, 1 = Uncertainty, 2 = Anxiety)
3. **Storage Layer**
    - Secure, scalable storage of processed text and corresponding model outputs
    - Metadata tagging for time-series analytics and longitudinal tracking
4. **Visualization and Reporting Dashboard**
    - Interactive dashboards showing:
        - Sentiment trends over time
        - Distribution of emotional categories
        - Emerging topics and discourse shifts
        - Alerts when anxiety spikes or hype drops
    - Designed for HR teams, policymakers, academic researchers, or industry analysts

8.3 Practical Use Cases

The deployed system can support several practical applications:

Workforce Sentiment Monitoring: Organizations can track how employees or industry communities express concerns about automation, layoffs, skill displacement, and technological disruption.

Policy and Labor Market Research: Policy institutions and labor economists can use the system to examine:

- Anxiety patterns during economic shocks

- Public response to major AI announcements
- Differences in discourse across demographic or occupational groups

Industry Trend Detection: AI and tech companies can identify:

- Peaks in hype-driven attention
- Negative sentiment clusters associated with new system releases
- Reactions to regulatory changes

Academic Research: The pipeline provides:

- An automated means of collecting and analyzing large-scale public discourse
- A replicable framework for studying socio-emotional responses to AI
- A baseline model that can be extended to multilingual or multimodal contexts

8.4 Model Maintenance and Continuous Improvement

To ensure long-term reliability, the deployment plan includes:

- **Periodic Model Retraining:** Linguistic patterns on Reddit evolve quickly; the model should be retrained quarterly or semi-annually.
- **Drift Detection:** Monitoring for shifts in vocabulary, sentiment distributions, or topic coherence.
- **Human-in-the-Loop Validation:** Periodic manual audits of model predictions to ensure label accuracy and prevent category drift.
- **Ethical Oversight:**
  Because the model works with public but potentially sensitive discourse, data collection must comply with platform guidelines and privacy norms.

8.5 Deployment Environment

A minimal deployment configuration could be implemented using:

- **Backend:** Python + FastAPI or Flask
- **ML Serving:**
  - scikit-learn models for TF-IDF classifiers
  - Transformers served via HuggingFace or ONNX Runtime
- **Data Pipeline:** Airflow, Prefect, or cron-based scheduling
- **Dashboard:** Streamlit, Plotly Dash, or Power BI
- **Cloud Providers:** AWS, GCP, or Azure for scalable storage and compute

This architecture is intentionally lightweight and cost-effective, making it feasible for non-profit organizations, academic labs, or small industry teams.

8.6 Limitations and Scalability Considerations

While the proposed deployment architecture demonstrates how the analytical pipeline can be operationalized, several limitations should be acknowledged from a deployment and scalability perspective. First, the current system is trained on Reddit data, which may reflect platform-specific discourse norms and overrepresent technology-oriented user populations. As a result, deployment to broader labor-monitoring contexts would require additional platform-specific adaptation.

Second, emotional category labels are derived using a hybrid rule-based framework rather than manual annotation. While this approach enables scalability, it introduces boundary ambiguity—particularly for the Uncertainty category—which may affect downstream interpretability in high-stakes applications. Incorporating human-in-the-loop validation or active learning could improve robustness in production settings.

Finally, the current deployment design focuses on document-level inference and does not model temporal dynamics or user-level trajectories. Future extensions could integrate longitudinal monitoring to track emotional shifts over time, enabling early detection of emerging labor anxieties or optimism trends. Despite these limitations, the modular design of the system supports incremental enhancement without requiring fundamental changes to the modeling pipeline.

8.7 Summary

The deployment plan translates the analytical insights of the project into a practical system capable of monitoring AI-related emotional discourse in real time. By integrating dynamic data ingestion, interpretability-focused analytics, and scalable serving infrastructure, the system provides a robust foundation for understanding how workers and online communities perceive AI's evolving role in labor markets.

This design not only satisfies the rubric requirement for describing practical application but also positions your project as a credible prototype for real-world sentiment and discourse monitoring tools.

9. Summary and Conclusion

This project set out to examine how individuals emotionally engage with artificial intelligence in relation to work, job security, and career futures by analyzing large-scale, real-world online discourse. Rather than

treating AI sentiment as a simple positive–negative dichotomy, the study framed emotional orientation as a more nuanced construct—distinguishing between **Career Anxiety**, **Future/Hype**, and **Uncertainty**—and investigated whether these orientations could be described, contextualized, and predicted using natural language processing techniques.

Using a custom-built Reddit corpus of approximately 10,000 posts and comments, the analysis demonstrated that discussions about AI and work are deeply emotional, linguistically structured, and thematically coherent. Preprocessing and exploratory data analysis revealed a corpus rich in employment-centered vocabulary, evaluative language, and forward-looking speculation. Sentiment analysis using VADER confirmed that emotional polarity varies systematically across discourse types, with anxiety-oriented posts exhibiting lower compound scores, hype-oriented posts showing strong positive polarity, and uncertainty occupying a meaningful intermediate space.

Topic modeling further enriched these findings by uncovering stable thematic patterns that align closely with emotional categories. Across classical models (LDA, NMF) and embedding-based approaches (Word2Vec, FastText, GloVe, BERT), consistent themes emerged: anxiety-related discourse centered on layoffs, burnout, managerial strain, and financial insecurity, while future-oriented discourse emphasized innovation, skill-building, career mobility, and optimism about technological progress. Importantly, uncertainty was not noise or misclassification error, but a distinct emotional stance characterized by exploration, hesitation, and mixed affect. These results illustrate that emotional tone is embedded not only in sentiment polarity but also in topic structure and lexical organization.

The supervised learning stage demonstrated that emotional orientation toward AI-related career discourse is **predictively learnable from text**. Among the evaluated models, linear classifiers—particularly **Linear Support Vector Machine (SVM)**—consistently outperformed probabilistic and non-linear approaches when applied to TF–IDF features. Linear SVM achieved the best balance across accuracy, precision, recall, and F1-score, confirming that margin-based linear decision boundaries are well suited for sparse, high-dimensional text data. Errors across models were systematic and concentrated around the boundary between anxiety and uncertainty, reflecting genuine semantic overlap rather than modeling failure.

Taken together, the findings support several broader conclusions. First, emotional responses to AI and work are neither random nor purely individual; they follow discernible linguistic and thematic patterns that can be studied computationally. Second, uncertainty plays a critical role in AI-related discourse and should not be collapsed into positive or negative sentiment. Third, model performance depends less on algorithmic complexity and more on alignment between representation, task framing, and data characteristics.

Beyond its technical contributions, this project highlights the value of integrating computational methods with socio-emotional inquiry. By examining how people talk about AI in their own words, the analysis provides insight into emerging labor anxieties, evolving expectations, and the narratives shaping public responses to technological change. As AI continues to transform work, such emotionally grounded analyses can complement economic and policy-focused approaches, offering a more human-centered understanding of technological impact.

In sum, this study demonstrates that natural language processing can be used not only to model text, but also to surface the emotional realities of technological transition—capturing fear, hope, and uncertainty as they unfold in real time

References

1. Acemoglu, D., & Restrepo, P. (2019). Artificial intelligence, automation, and work. *Journal of Economic Perspectives, 33*(2), 197–236. https://ideas.repec.org/p/nbr/nberwo/24196.html
2. Autor, D. H. (2015). Why are there still so many jobs? The history and future of workplace automation. *Journal of Economic Perspectives, 29*(3), 3–30. https://www.researchgate.net/publication/282320407_Why_Are_There_Still_So_Many_Jobs_The_History_and_Future_of_Workplace_Automation
3. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*, 993–1022. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf
4. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 4171–4186. https://aclanthology.org/N19-1423/
5. Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change, 114*, 254–280. https://www.sciencedirect.com/science/article/pii/S0040162516302244
6. Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, 216–225. https://ojs.aaai.org/index.php/ICWSM/article/view/14550
7. Jurafsky, D., & Martin, J. H. (2025). *Speech and language processing* (3rd ed., draft). Stanford University. https://web.stanford.edu/~jurafsky/slp3/
8. Landauer, T. K., Foltz, P. W., & Laham, D. (1998).An introduction to latent semantic analysis. *Discourse Processes, 25*(2–3), 259–284. https://www.researchgate.net/publication/200045222_An_Introduction_to_Latent_Semantic_Analysis
9. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *Proceedings of the International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1301.3781
10. Pennington, J., Socher, R., & Manning, C. D. (2014).GloVe: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://aclanthology.org/D14-1162/
11. Rehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. https://www.researchgate.net/publication/255820377_Software_Framework_for_Topic_Modelling_with_Large_Corpora
12. Salton, G., & Buckley, C. (1988).Term-weighting approaches in automatic text retrieval. *Information Processing & Management,* *https://www.sciencedirect.com/science/article/pii/0306457388900210*