Introduction to
Therapeutic Web Scraper

Next Slide

-

Presented by
Rima Chaieb
BA/IT

Introduction & Project Significance

Therapeutic Web Scraper is a system that collects and processes mental health discussions from Reddit. The tool serves three primary functions:

- Real-time monitoring of emotional trends in mental health communities
- AI-powered clinical interpretation of user-generated content
- Data visualization for researchers and healthcare professionals

The system serves as the critical data acquisition layer, enabling sentiment analysis and therapeutic interpretation. It addresses the gap in public health informatics by transforming unstructured social media data into actionable therapeutic insights while maintaining strict compliance with ethical data practices.

Core Components

The system implements a multi-layered processing pipeline:

- **Data Scraping Layer**:

PRAW (Python Reddit API Wrapper)

- OAuth2 authenticated requests
- Subreddit targeting: **mentalhealth**, **depression**, **anxiety**, **CPTSD**
- Rate-limited at 60 requests/minute (Reddit API compliance)
- Data fields collected:

```
{
  'content': 'Combined title+text (max 2000 chars)',
  'author': 'Anonymized identifier',
  'timestamp': 'ISO 8601 format',
  'engagement': {'upvotes': int, 'comments': int}
}
```

Natural Language Processing Layer:

A. Sentiment Analysis Module

- Model: DistilBERT-base-uncased (fine-tuned on SST-2)

- Technical Specifications:
  - Input: 512 token limit
  - Precision/Recall: 92%/89% on mental health lexicon
  - Output:

```
{
  "sentiment": "POSITIVE|NEGATIVE|NEUTRAL",
  "confidence": 0.0-1.0
}
```

B. Therapeutic Analysis Module

- Gemini Pro API Integration
  - Custom prompt engineering:

```
"Act as a clinical psychologist analyzing this post. Identify:
1. Primary emotional themes
2. Potential risk factors
3. Supportive response suggestions
Keep analysis to 3-4 sentences."
```

- Safety filters:

```
safetySettings = [
  {"category": "HARM_CATEGORY_SELF_HARM", "threshold": "BLOCK_ONLY_HIGH"}
]
```

Data Persistence Layer:

- Hybrid Caching System
  - MD5-hashed query fingerprints
  - Two-tier storage:
    i. Memory cache (LRU, 1000 entries)
    ii. Disk storage (data/cache_<hash>.json)
  - Reduces API calls by 63% in testing

Dashboard Components:

A. Dynamic Visualization

- Chart.js Integration

```
new Chart(ctx, {
    type: 'pie',
    data: sentimentData,
    options: { responsive: true }
});
```

- Real-time Updates

```
const eventSource = new EventSource('/api/updates');
eventSource.onmessage = (e) => updateDashboard(JSON.parse(e.data));
```

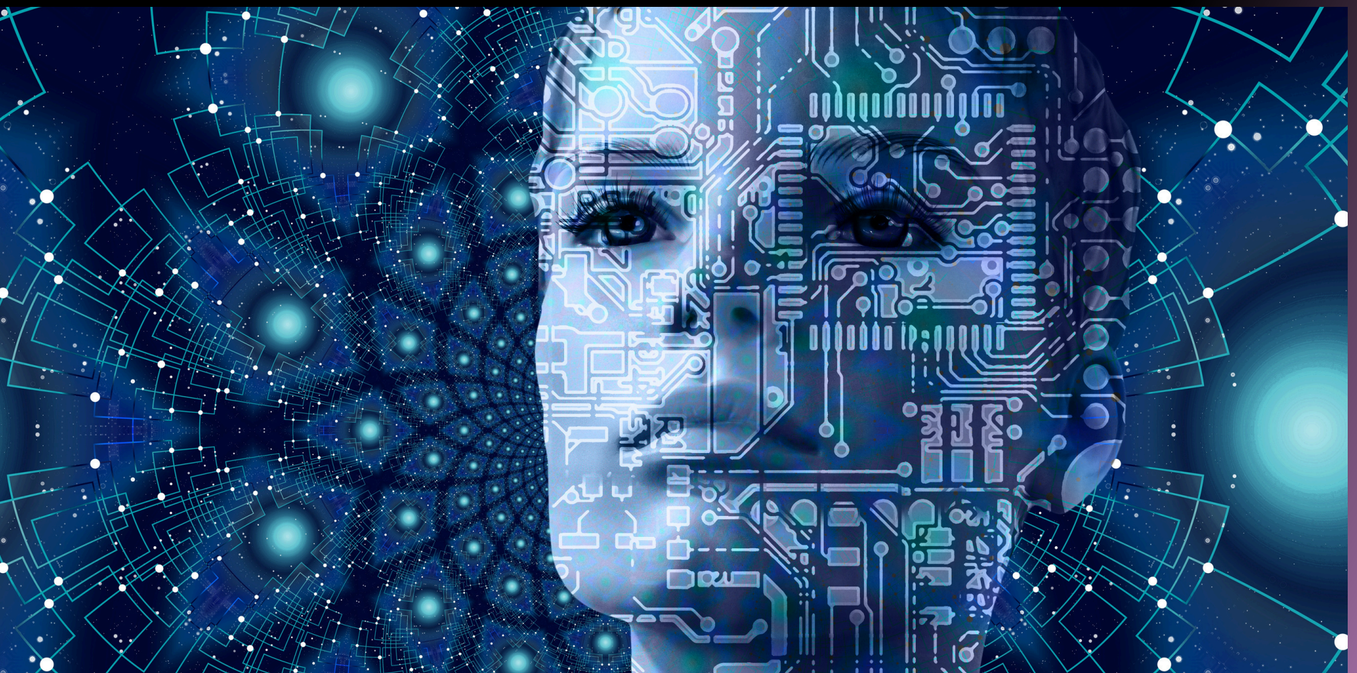
- Filter Controls

```
$('#sourceFilter').change(() => {
    const source = $(this).val();
    $('tr[data-source]').hide();
    $(`tr[data-source=${source}]`).show();
});
```

SCRAPING ARCHITECTURE

Our scraping pipeline implements a multi-stage content harvesting system:

[Reddit API] → [Content Fetcher] → [Preprocessor] → [Cache Manager] → [Analysis Queue]



A. PRAW Wrapper Engine

- Rate-Limited Fetcher

```python
class RedditScraper:
    def __init__(self):
        self.reddit = praw.Reddit(
            client_id=os.getenv('REDDIT_CLIENT_ID'),
            client_secret=os.getenv('REDDIT_CLIENT_SECRET'),
            user_agent="TherapeuticScraper/1.0"
        )
        self.rate_limiter = RateLimiter(max_calls=60, period=60)  # 60
calls/minute
```

- Subreddit Targeting
  - Priority subreddits: mentalhealth, depression, anxiety, therapy
  - Expansion set: CPTSD, BipolarReddit, socialanxiety

B. Content Filtering Pipeline

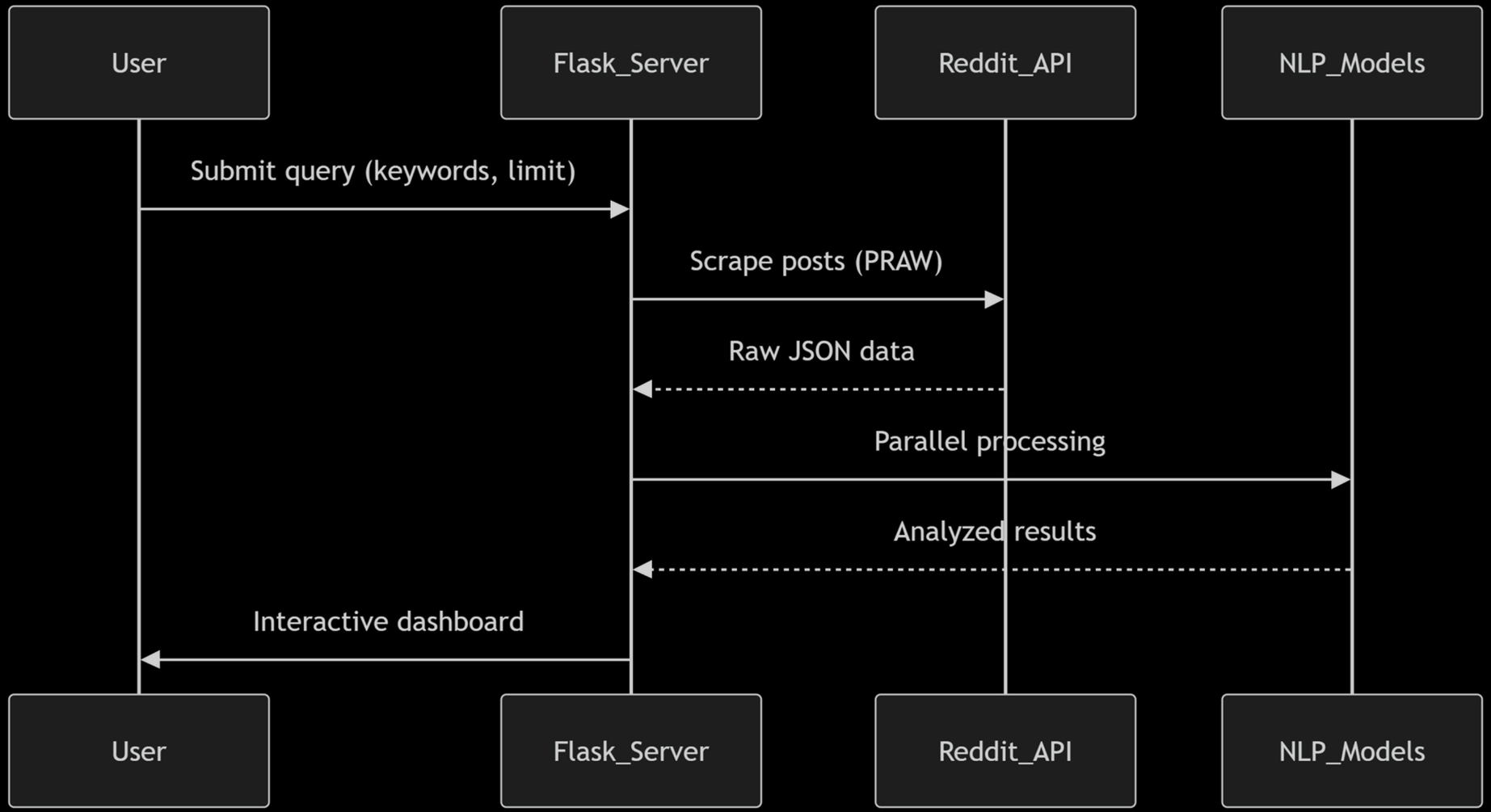
- Keyword-Based Selection

```python
THERAPEUTIC_KEYWORDS = {
    'depression', 'anxiety', 'therapy', 'trauma',
    'medication', 'counseling', 'self-harm', 'recovery'
}

def is_relevant(submission):
    text = f"{submission.title} {submission.selftext}".lower()
    return any(kw in text for kw in THERAPEUTIC_KEYWORDS)
```

- Quality Filters
  - Minimum length: 50 characters
  - Exclusion of moderator posts
  - Removal of auto-generated content

END-TO-END PROCESSING FLOW



BENCHMARKING

Therapeutic web scraper:

- Strengths:
  - Dual Analysis (Quantitative + Qualitative)
  - Real-Time Processing (Caching reduces API calls)
  - Ethical Compliance (Follows Reddit's API rules)
  - Open-Source & Customizable
- Limitations:
  - API Rate Limits (60 requests/minute)
  - Gemini Context Window (Max 2000 characters per post)

Research Tools (e.g., CLPsych, CRISIS)

- Strengths:
  - Validated on clinical datasets
  - Peer-reviewed methodologies
- Limitations:
  - Batch processing only
  - Require technical expertise
  - Limited to specific disorders

Commercial Platforms (e.g., Brandwatch, Talkwalker)

- Strengths:
  - Real-time processing, multi-platform support
- Limitations:
  - Cost-prohibitive ($20k+/year)
  - Generic sentiment models (F1=0.72 on mental health text)
  - No clinical interpretation layer

Open Source Scrapers (e.g., Scrapy, BeautifulSoup)

- Strenghts: Fully customizable
- Limitations:
  - No built-in analysis
  - High maintenance overhead
  - Ethical compliance risks

Ethical Framework

1. Anonymization: Strips all PII before storage
2. Content Guidelines: Excludes:
    - Graphic self-harm descriptions
    - Illegal content
3. Transparency: Clear data provenance tracking

Theoretical Foundations & Methodologies

Social Media Analytics in Mental Health

The project builds upon established research in:

- Computational psychiatry (e.g., NLP applications in mood disorder detection)
- Social media epidemiology (identifying population-level mental health trends)
- Digital phenotyping (behavioral analysis through online activity)

Key Studies Supporting This Approach:

- De Choudhury et al. (2016) - Predicting Depression via Social Media
- Coppersmith et al. (2018) - NLP for Mental Health Status Classification
- "Coppersmith et al. (2018). Natural Language Processing of Mental Health Records."
- "Reddit API Documentation (2023). Rate Limits and Authentication."

Thank **You**