Mental disorder detection from social media data

Tags

https://prod-files-secure.s3.us-west-2.amazonaws.com/dda62111-67f9-4bc2-8d24-91d9290e5fa9/f35e757b-7ca3-484b-80b8-762927d8aa7e/Slides_compressed.pdf

Data Collection

Direct Participant Data Collection

- **Questionnaires:** Surveys delivered to participants to gather specific information about their experiences, conditions, or opinions.
- **Electronic Health Records (EHR):** Extracting relevant data from participants' medical records with their consent.

Social Media Data Aggregation

- **Keyword Search:** Identifying relevant data by searching public social media posts for specific keywords or phrases like "I was diagnosed with [condition name]".
- **Annotation:** Manually tagging or classifying extracted data to categorize and analyze it effectively.

Benefits and Considerations

- **Direct Participant Data**

Pros: Controlled data quality, targeted information gathering, participant consent and ethical considerations.

Cons: Potential for response bias, time and cost required to recruit and survey participants.

- **Social Media Data**

Pros: Large volume of data available, insights into real-world language and behavior, potential for cost-effectiveness.

Cons: Privacy concerns, data quality challenges due to ambiguity and misinformation, ethical considerations regarding non-consensual data collection.

Data Exploration & Preprocessing

Feature Extraction:

- This involves transforming raw text data into numerical representations that computers can understand.
- Techniques like **bag-of-words** (as shown in the example) analyze word frequency and create vectors for each document.
- Other tools like **LIWC** analyze linguistic features like pronouns and emotional tones.