Contents

Social Network Analytics - User Guide

Table of Contents

Introduction

Welcome to Social Network Analytics! This tool helps you extract and visualize social networks from social media data using various extraction methods including Named Entity Recognition (NER), hashtags, mentions, keywords, and more.

What Can You Do?

- **Extract networks** from CSV or NDJSON files
- **Multiple extraction methods**: entities, hashtags, mentions, keywords, domains, or exact matches
- **Multilingual support**: Works with 105+ languages for NER
- **Interactive visualization**: Explore your network with Force Atlas 2 layout
- **Export formats**: GEXF (Gephi), GraphML, JSON, CSV, and more

---

Getting Started

Launching the Application

1. Open your terminal/command prompt

2. Navigate to the application directory
3. Run: `streamlit run src/cli/app.py`
4. Your browser will open automatically to the application

What You'll See

The application has a sidebar (left) for configuration and a main area (center/right) for data upload, processing, and results.

Data Upload

Step 1: Upload Your Data

Supported File Formats: - **CSV**: Comma-separated values - **NDJSON/JSONL**: Newline-delimited JSON

File Requirements: - Must contain at least two columns: one for author/user and one for text/content - Text column should contain the posts, comments, or messages to analyze

Step 2: Preview Your Data

After uploading, you'll see: - **Data preview**: First 10 rows of your data - **File information**: Size, total rows, number of columns - **Column list**: All available columns in your file

Step 3: Select Columns

Author Column: - The column containing usernames, author names, or user IDs - Examples: `username`, `author`, `user_id`, `unique_id`

Text Column: - The column containing the text to analyze - Examples: `text`, `content`, `message`, `post`, `body`

Smart Detection: The app automatically tries to detect these columns based on common naming patterns. If the selection is wrong, simply change it using the dropdown menus.

Step 4: Metadata Columns (Optional)

You can attach additional data to your network:

Node Metadata: - Attached to author nodes - If multiple posts have different values, only the first value is kept - Examples: follower count, location, verified status

Edge Metadata: - Attached to connections between authors and entities - Multiple unique values are stored as a list - Examples: post text, timestamps, engagement metrics

Tip: You can include the text and author columns as metadata to preserve original post content on edges!

Extraction Methods

Choose how you want to extract items from your text to build the network:

1. NER (Named Entities) - Default

What it does: Uses AI to identify people, places, and organizations in text

Best for: - News articles - Social media discussions about real-world entities - Multilingual content

Configuration: - **Model**: Choose between two multilingual models - **Entity Types**: Select persons, locations, and/or organizations - **Confidence**: Higher = more accurate but fewer results (default: 85%) - **Batch Size**: Higher = faster processing with more memory

Example: - Input: "Joe Biden met with Emmanuel Macron in Paris" - Output: Joe Biden (PERSON), Emmanuel Macron (PERSON), Paris (LOCATION)

2. Hashtags

What it does: Extracts hashtags from text

Best for: - Twitter/X data - Instagram posts - TikTok content

Configuration: - **Normalize case**: Convert #Python to #python

Example: - Input: "Love this #Python tutorial! #Coding #DataScience" - Output: python, coding, datascience

3. Mentions (@username)

What it does: Extracts user mentions from text

Best for: - Twitter/X conversations - Instagram comments - Any platform with @mentions

Configuration: - **Normalize case**: Convert @User to @user

Example: - Input: "Great work @john_doe and @jane_smith!" - Output: john_doe, jane_smith

Note: The @ symbol is removed to allow matching between authors and mentioned users.

4. URL Domains

What it does: Extracts domain names from URLs

Best for: - Link sharing analysis - News source tracking - Website reference networks

Configuration: - **Strip 'www.' prefix**: Convert www.example.com to example.com

Example: - Input: "Check out https://www.nytimes.com/article and http://bbc.co.uk/news" - Output: nytimes.com, bbc.co.uk

5. Keywords (RAKE)

What it does: Uses RAKE algorithm with TF-IDF weighting to extract important keywords

Best for: - Topic analysis - Content summaries - Thematic networks

Configuration: - **Min/Max keywords per author**: How many keywords to extract (default: 5-20) - **Language**: Choose language for stopword filtering - **Max phrase length**: Maximum words in a keyword phrase (default: 3)

Advanced Features: - TF-IDF weighting boosts distinctive keywords - Filters out common words, URLs, hashtags, mentions - Comprehensive stopword lists for Danish and English

Example: - Input: Multiple posts about climate change - Output: climate change, renewable energy, carbon emissions, global warming

6. Exact Match

What it does: Uses the text value as-is without extraction

Best for: - Pre-processed data - Category labels - Simple text matching

Example: - Input: "sports" - Output: sports

Configuration Options

Basic Settings (All Methods)

Chunk Size: - Number of rows to process at once - Higher = faster but more memory - Default: 10,000 rows

NER-Specific Settings

NER Models: 1. **Davlan/xlm-roberta-base-ner-hrl** (default) - Multilingual: 10 high-resource languages - Fast and accurate

2. **Babelscape/wikineural-multilingual-ner**
   - Multilingual: 105 languages
   - Slower but broader language support

Entity Types: - **Persons (PER)**: Names of people - **Locations (LOC)**: Cities, countries, regions - **Organizations (ORG)**: Companies, institutions

Confidence Threshold: - Minimum confidence score (0.5 - 1.0) - Higher = more precise, fewer results - Lower = more results, some false positives - Default: 0.85 (85%)

Batch Size: - Number of texts to process together - Higher = faster with GPU - Range: 8-128 - Default: 32

Advanced Options

Cache Settings (NER only): - **Enable NER Cache**: Speeds up reprocessing - **Clear Cache**: Forces fresh extraction

Language Detection (NER only): - Automatically detects language of each post - Useful for multilingual datasets

Author-to-Author Edges: - Creates edges when authors mention each other - Only works with mention extraction

Entity Deduplication: - Merges similar entities (case-insensitive) - Recommended: Keep enabled

Entity Linking (NER only - Experimental): - Links entities to Wikipedia/Wikidata - Helps merge cross-language variants - Example: “København” = “Copenhagen” = “Copenhague” - **Warning**: Slower processing and may have issues

Visualization Quality: - Force Atlas iterations (50-200) - Higher = better layout, slower rendering - Default: 100

Processing Your Data

Click “Start Processing”

Once you’ve configured everything, click the “ **Start Processing”** button.

What Happens During Processing

1. **Initialization**: Model loading (first time only)
2. **Progress Tracking**:
   - Progress bar showing completion
   - Estimated time remaining

    - Current status messages
3. **Completion**: Success message with statistics

Processing Times

Factors: - File size (number of rows) - Extraction method (NER is slowest) - Batch size and chunk size - Hardware (GPU vs CPU)

Typical Speeds: - **NER with GPU**: 50-100 posts/second - **NER with CPU**: 10-30 posts/second - **Hashtags/Mentions**: 1000+ posts/second - **Keywords (RAKE)**: 100-500 posts/second

Troubleshooting

"Out of Memory" Error: - Reduce batch size - Reduce chunk size - Close other applications

"No entities extracted": - Check text column has content - Lower confidence threshold (NER) - Try different extraction method

Very slow processing: - Ensure GPU is being used (check startup messages) - Increase batch size if using GPU - Consider using simpler extraction method

---

Understanding Results

Statistics Overview

After processing, you'll see several metrics:

Network Metrics: - **Total Nodes**: Authors + Entities - **Total Edges**: Connections between authors and entities - **Authors**: Number of unique authors - **Entities**: Number of unique extracted items - **Density**: How connected the network is (0-1)

Entity Breakdown (NER only): - **Persons**: Individual people identified - **Locations**: Places, cities, countries - **Organizations**: Companies, institutions

Processing Details: - **Posts Processed**: How many rows were analyzed - **Chunks Processed**: Number of batches processed - **Entities Extracted**: Total items found (before deduplication) - **Processing Speed**: Posts per second

Top Mentioned Entities

A table showing the most frequently mentioned entities:

Columns: - **Entity**: The name of the entity - **Mentions**: How many times it was mentioned - **Type**: Category (person, location, organization, etc.)

With Entity Linking: - **Wikidata ID**: Unique identifier (e.g., Q1748) - **Wikipedia**: Link to Wikipedia page

Color Coding: - Blue: Persons - Orange: Locations - Purple: Organizations

---

Network Visualization

Interactive Force Atlas 2 Visualization

Your network is displayed using an interactive, physics-based layout.

Visualization Controls

Giant Component Only: - Toggle to show only the largest connected group - Helps focus on the main network - Useful for large networks with isolated clusters

Performance Limits: - Networks over 1,000 nodes show top 500 most connected - This ensures smooth performance

How to Interact

Mouse Controls: - **Hover**: See node details (name, type, connections) - **Scroll**: Zoom in/out - **Click + Drag**: Pan around the network

Right-Side Controls: - **Play/Pause**: Start/stop the physics simulation - **Settings**: Adjust layout parameters in real-time

Understanding the Visualization

Node Colors: - **Blue**: Authors (people posting) - **Orange**: Persons (mentioned entities) - **Green**: Locations - **Red**: Organizations - **Other colors**: For other extraction methods

Node Size: - Larger = More connections - Smaller = Fewer connections

Edges (Connections): - Line thickness = Connection strength (how many times mentioned) - Direction: Author → Entity

Reading the Network

Clusters: - Groups of closely connected nodes - May represent topics, communities, or themes

Central Nodes: - Large nodes in the middle - These entities/authors are most connected

Isolates: - Nodes far from others or alone - Less connected to the main discussion

---

Downloading Results

Available Formats

Primary Format:

GEXF (for Gephi): - Best for network analysis in Gephi - Preserves all network attributes - Recommended format

Additional Formats:

GraphML: - Compatible with many network tools - Good alternative to GEXF

JSON (D3.js): - For web visualizations - Used in D3.js and other JavaScript libraries

Edge List CSV: - Simple format: source, target, weight - Easy to import into R, Python, Excel

Statistics (JSON): - All network metrics in JSON format - For further analysis or reporting

Using Exported Files

With Gephi: 1. Download GEXF file 2. Open Gephi 3. File → Open → Select your .gexf file 4. Run layout algorithms (Force Atlas 2, etc.) 5. Customize appearance and export visualizations

With R/Python: 1. Download Edge List CSV 2. Import into network library (igraph, networkx, etc.) 3. Perform statistical analysis

For Reports: 1. Download Statistics JSON 2. Parse the data 3. Create custom tables and charts

Tips & Best Practices

Data Preparation

1. Clean Your Data First: - Remove empty rows - Ensure text column has content - Check for encoding issues

2. Sample Large Datasets: - Test with a small sample first - Verify extraction method works correctly - Then process full dataset

3. Choose the Right Extraction Method: - **News/Articles**: Use NER - **Twitter/Social Media**: Use hashtags or mentions - **Topic Analysis**: Use keywords - **Link Sharing**: Use domains

Optimization

1. Speed Up Processing: - Use GPU if available - Increase batch size (with GPU) - Enable caching (NER) - Use simpler extraction methods

2. Improve Quality: - Adjust confidence threshold (NER) - Enable entity linking for cross-language (NER) - Use entity deduplication - Fine-tune keyword settings (RAKE)

3. Memory Management: - Reduce chunk size if running out of memory - Close other applications - Process in batches for very large files

Network Analysis

1. Start Simple: - Use giant component filter to focus on main network - Look at top mentioned entities first - Identify obvious clusters

2. Iterative Analysis: - Try different extraction methods - Adjust confidence thresholds - Compare results

3. Export for Detailed Analysis: - Use Gephi for advanced layouts - Calculate centrality measures - Identify communities

Common Use Cases

1. Topic Discovery: - Use keyword extraction (RAKE) - Look for clusters in visualization - Examine top entities

2. Influence Analysis: - Use mention extraction - Node size = influence - Identify key opinion leaders

3. Information Flow: - Use domain extraction - See which sources are shared - Identify information brokers

4. Multilingual Analysis: - Use NER with entity linking - Merges entities across languages - Unified global view

Troubleshooting Common Issues

Problem: No results/empty network - Check text column has content - Verify correct columns selected - Lower confidence threshold (NER) - Try different extraction method

Problem: Too many irrelevant entities - Increase confidence threshold (NER) - Enable keyword filters (RAKE) - Adjust min/max keywords - Use more specific extraction method

Problem: Slow processing - Reduce batch size (if CPU) - Increase batch size (if GPU) - Enable caching - Use simpler extraction method

Problem: Entities not merging - Enable entity deduplication - Enable entity linking (experimental) - Use normalize case option - Manually merge in Gephi after export

Frequently Asked Questions

Q: What file formats are supported? A: CSV and NDJSON (newline-delimited JSON) files.

Q: How large can my file be? A: The tool can handle files with millions of rows, but processing time increases. Start with samples for large files.

Q: Do I need a GPU? A: No, but it's highly recommended for NER extraction. CPU works but is 3-5x slower.

Q: Can I process multiple languages? A: Yes! NER supports 105 languages. Choose appropriate model and enable entity linking.

Q: How do I cite this tool? A: Documentation for citation will be provided separately.

Q: Where is my data stored? A: All processing happens locally. Data is cached temporarily in `./cache/` directory.

Q: Can I customize the extraction? A: Yes, through configuration options. For advanced customization, see the technical documentation.

Q: Why are some entities not extracted? A: Could be due to confidence threshold, text quality, or model limitations. Try adjusting settings.

Q: How do I get the best visualization? A: Export to GEXF and use Gephi for professional-quality network visualizations with full control.

Q: Can I process the same data multiple times? A: Yes! NER results are cached, making reprocessing very fast.

Getting Help

Issues or Questions: - Check this user guide first - Review the troubleshooting section - Experiment with different settings on small samples

For Technical Support: - GitHub Issues: [Repository link will be provided] - Include: error messages, file format, settings used

---

Version Information

Current Version: 2.0 **Last Updated:** December 2025

Happy Network Analysis!