Product Requirements Document (PRD)

AI-Powered Press Clipping Agency Application

1. Executive Summary

This document outlines the requirements for an AI-powered Press Clipping Agency Application that automatically monitors, scrapes, and analyzes news media portals and associated social media accounts for specific search terms. The system will deliver comprehensive daily reports via email in Markdown format.

2. Product Overview

2.1 Title of the Application

The title of the application is “Press Clipping AI Agent”, and this is for now only a working name. Once we are close to the end of developing the app, the name shall be discussed.

2.2 Product Vision

An intelligent automation system that replaces traditional press clipping services by leveraging AI agents to conduct deep research across news media platforms and social media channels, delivering curated insights on specified topics.

2.3 The Problem it Solves

- The user shall have an overview of news published about specific persons and events related to the person on news web portals, both local and national, as well as a basic report when particular keywords are mentioned on Facebook, Twitter/X, X, Instagram, and YouTube.
- The report, presented in tabular form, includes columns such as the rolling number of the event, date and time, a short description, the location where the keyword was mentioned, and the sources where the keyword appeared.
- The report shall be received via email and made accessible at any time.
- The app shall create a value such that, once an email is received, the user of the app will be communicated to the Public Relations Officer(s), who will take care of and respond to ensure that no further damage is caused.

2.4 Target Users

- Communications team or City of Niš PR professional and city mayor itself
- Public relations professionals
- Political organizations (optional)
- Corporate monitoring departments (optional)
- Research institutions (optional)

---

3. Core Features & Requirements

3.1 Functional Requirements

FR-1: Data Source Management

- System shall read and parse **`mediaportalsfull.md`** file containing prioritized media outlets
- File structure includes: News media portals with URLs and social media accounts
- Priority system: Column 1 = highest priority, ascending row numbers = descending priority

FR-2: Search & Scraping Capabilities

- Conduct deep research on specified search terms
- Scrape data from:
  - Primarily from news media portal URLs which are listed in the first coloune
  - Secondarily, search for the associated social media accounts (Twitter, Facebook, Instagram, LinkedIn, etc.)
- Extract:
  - Article/post titles
  - Brief descriptions/summaries
  - Direct URLs to source content
  - Publication timestamps
  - Author information (when available)

FR-3: Time-Based Filtering

- **Primary time-frame**: Last 24 hours (for daily 06:00 AM run)
- **Secondary time-frame**: Previous 72 hours (for context/comparison) only on Mondays
- Timestamp all results with publication date/time

FR-4: Automated Scheduling

- **Daily execution**: Automatically activate at 06:00 AM
- **Manual activation**: On-demand execution via API call or command
- Post-06:00 AM activation focuses on current day's data from 06:00 AM on the current day

FR-5: Report Generation

- Generate email-ready Markdown format reports
- Include:
  - Executive summary
  - Findings organized by media source (priority-ordered)
  - Search term analysis
  - Time-frame segmentation (24h vs 72h(optional))
  - Direct links to all sources
  - Metadata (search date, terms used, sources scanned)

FR-6: Email Delivery

- Send completed reports to designated email address(es)
- Subject line format: **`[Press Clipping Report] - [Date] - [Search Terms]`**
- Support multiple recipient addresses
- Include failure notifications

3.2 Technical Requirements

TR-1: Platform & Framework

- Built on **You.com platform**
- Based on **Langchain Academy Deep Research Agent** repository
- Repository URL: https://github.com/langchain-ai/deep_research_from_scratch

TR-2: Architecture Components

- Research agent module
- Web scraping module using tools such as
  - **News Media Web Portals**
  - **Tweeter/X**: Free API Tier
  - **Facebook**:
    - Official API Tier – 200 API calls per hour
    - Unofficial scrappers like **`facebook-scraper`** do not have official rate limits
  - **Instagram**:
    - Official API Free Tier 200 API calls
    - Unofficial scrapers like **`instaloader`** are not rate limited
  - **YouTube**: Limit is up to 10,000 quota units per day
- Social media API integration module
- Data processing & analysis module
- Report generation module
- Email delivery module
- Scheduling service

TR-3: Data Storage

- **`mediaportalsfull.md`** file stored in application repository
- consider **`PostgreSQL`** database with login and history data
- Configuration for search terms
- Historical report archive (optional)
- Error logs and execution logs

TR-4: Integration Requirements

- Social media API access (Twitter/X, Facebook, Instagram, LinkedIn)
- Email service integration (SMTP/SendGrid/AWS SES)
- Web scraping infrastructure with **rate limiting**
- Proxy support for geographic restrictions

TR-5: Performance Requirements

- Complete research cycle within 30-45 minutes
- Handle minimum 50 media sources simultaneously
- Process minimum 10 search terms per execution
- 99% up-time for scheduled runs

---

4. System Prompt Architecture

4.1 Complete System Prompt Structure

```
## ROLE
You are an AI Press Clipping Research Agent specialized in media monitoring and
intelligence gathering. You function as an automated journalist and analyst,
tasked with comprehensively tracking news coverage and social media discussions
across prioritized media sources.

## TASK
Your primary tasks are:
1. Monitor, fetch and scrape content from specified news media portals and their
associated social media accounts
2. **Based on the date and time posted**, search for predefined terms across all
sources within specified time-frames of the last 24h and 72h only on Mondays
3. Extract relevant articles, posts, and discussions containing the search terms
4. Analyze and summarize findings with context and relevance scoring
5. Compile a structured, email-ready report in Markdown format
6. Deliver the report via email to designated recipients

## CONTEXT
- **Execution Schedule**: You activate automatically at 06:00 AM daily or can be
triggered manually
- **Data Source**: `mediaportalsfull.md` file, `Product-Requirements-
Document.pdf` file, and web sites links file containing prioritized tables of:
  - The news media portals (URLs + social media accounts)
- **Priority System**: Sources are ranked by column number (1 = highest) and row
position (lower numbers = higher priority)
- **Search Time-frames**:
  - Primary: The Last 24 hours (main focus for daily reports)
  - Secondary: Previous 72 hours only on Mondays(contextual reference)
- **Search Terms**: Configurable list of keywords, phrases, organization names,
or topics to monitor, adding option to be hard-coded or edited in the search bar
- **Output Destination**: Email delivery to configured recipient list

## REASONING
Your decision-making process should follow this logic:

1. **Source Prioritization**: Begin with highest-priority sources (column 1, row
1) and work systematically through the list
2. **Relevance Filtering**: Include content only when search terms appear in:
   - Headlines/titles
   - Article body/post content
   - Tags or categories
   - Author attributions (when relevant)
3. **Context Extraction**: For each match, determine:
   - Is this breaking news or ongoing coverage?
   - What is the sentiment (positive/neutral/negative)?
   - Is this primary reporting or commentary?
   - What is the source credibility level?
4. **Deduplication**: Identify and consolidate duplicate stories across sources
5. **Summarization**: Provide 2-3 sentence summaries that capture:
   - The main point
   - Key stakeholders or entities
   - Actionable implications

## OUTPUT
Generate a Markdown-formatted report in **professional Serbian language** using
**Latin characters** with the following structure:

### Report Header
- Report Title: "Press Clipping Report - [Date]"
- Generation timestamp
- Search terms used
- Time-frame covered
- Number of sources scanned

### Executive Summary
- Total items found
- Key trending topics
- Notable developments
- Alert-worthy items (crisis indicators, viral content)

### Findings by Source (Priority Order)
For each media source:
#### [Media Source Name]
- **Article/Post Title**: [Title with hyperlink]
- **Published**: [Date/Time]
- **Summary**: [2-3 sentence description]
- **URL**: [Direct link]
- **Relevance**: [High/Medium/Low]
- **Time-frame**: [24h or 72h]

### Findings by Search Term
Organize all results grouped by search term for cross-reference

### Social Media Highlights
- Top trending posts/threads
- Viral content related to search terms
- Engagement metrics (likes, shares, comments)

### Appendix
- Sources successfully scanned
- Sources with errors/unavailable
- API rate limits hit
- Processing time

## STOPPING
You should conclude your research and generate the report when:

1. **Completion Criteria Met**:
   - All sources in `mediaportalsfull.md` have been scanned
   - All search terms have been processed
   - Both time-frames (24h and 72h on Mondays) have been searched

2. **Time Limit Reached**:
   - Maximum execution time of 45 minutes exceeded
   - In this case, prioritize completed high-priority sources

3. **Critical Error Encountered**:
   - Email service unavailable (queue for retry)
   - More than 50% of sources inaccessible
   - Corruption in `mediaportalsfull.md` file

4. **Manual Override**:
   - Stop command issued by administrator
   - Emergency shutdown triggered

Upon stopping, always:
- Compile partial results if incomplete
- Log execution summary
- Note any errors or incomplete sections
- Send the report (or error notification) via email
- Reset state for next execution
```

5. Report Data Schema

5.1 mediaportalsfull.md Structure

```
# Media Portals Configuration

## News Media Overview Report

| Priority | Media Name | URL | Twitter | Facebook | Instagram | YouTube |
|----------|------------|-----|---------|----------|-----------|----------|
| 1 | [Media Name 1] | https://... | @handle | /page | @handle | /company |
| 2 | [Media Name 2] | https://... | @handle | /page | @handle | /company |
| ... | ... | ... | ... | ... | ... | ... |
```

6. User Stories

US-1: As a PR professional, I want to receive a daily digest of all media mentions at 6 AM so I can brief my team before the workday begins.

US-2: As a political campaign manager, I want to track mentions of our candidate and opponents across news and social media to inform our daily strategy.

US-3: As a crisis communications lead, I want to manually trigger the agent when breaking news occurs to get immediate comprehensive coverage.

US-4: As a media analyst, I want results organized by priority and time-frame so I can quickly identify urgent items.

7. Non-Functional Requirements

NFR-1: Reliability

- 99% successful daily execution rate
- Automatic retry logic for failed sources
- Graceful degradation if sources are unavailable

NFR-2: Security

- Secure storage of API keys and credentials
- Encrypted email transmission
- Rate limiting compliance with platform ToS

NFR-3: Scalability

- Support for 100+ media sources
- Handle 20+ search terms simultaneously
- Process 1000+ results per execution

NFR-4: Maintainability

- Easy update of **`mediaportalsfull.md`** without code changes
- Configurable search terms via **`config`** file
- Comprehensive logging and error tracking

NFR-5: Usability

- Markdown report readable in any email client
- Clear prioritization and organization
- Mobile-friendly formatting

8. Implementation Phases

Phase 1: Core Infrastructure (Weeks 1-2)

- Set up You.com platform environment
- Clone and adapt Langchain Deep Research Agent
- Implement `mediaportalsfull.md` parser
- Basic web scraping functionality

Phase 2: Social Media Integration (Weeks 3-4)

- Integrate social media **APIs**
- Implement authentication and rate limiting
- Test data extraction across platforms

Phase 3: Report Generation (Week 5)

- Build Markdown report generator
- Implement email delivery system
- Create report templates

Phase 4: Scheduling & Automation (Week 6)

- Implement cron/scheduler for 06:00 AM execution
- Add manual trigger mechanism
- Set up monitoring and alerts

Phase 5: Testing & Refinement (Weeks 7-8)

- End-to-end testing
- Performance optimization
- Bug fixes and refinements

---

9. Success Metrics

- **Execution Success Rate**: >99% of scheduled runs complete successfully
- **Coverage Completeness**: >95% of configured sources scanned per run
- **Delivery Time**: Reports delivered within 5 minutes of completion
- **False Positive Rate**: <10% irrelevant results included
- **User Satisfaction**: >4.5/5 rating on report usefulness

---

10. Risks & Mitigation

Risk	Impact	Probability	Mitigation
Social media API changes	High	Medium	Maintain multiple data sources, implement adapter pattern

Risk	Impact	Probability	Mitigation
Rate limiting blocking	Medium	High	Implement respectful delays, use proxies, cache results
Email delivery failures	High	Low	Queue-based retry, multiple email service providers
Website structure changes	Medium	High	Implement robust parsing with fallbacks
Overwhelming data volume	Medium	Medium	Implement smart filtering, relevance scoring

11. Dependencies

- You.com platform access
- Langchain Academy Deep Research Agent repository
- Social media API credentials (Twitter, Facebook, Instagram, LinkedIn)
- Email service provider account
- Web scraping infrastructure (proxies, user agents)
- Cron/scheduler service

12. Appendix

A. Sample Search Terms Configuration

```
search_terms:
  - "Dragoslav Pavlović"
  - "Gradonačelnik"
  - "Gradonačelnik Niša"
  - "Gradonačelnik grad aNiša"
  - ""
```

B. Sample Email Report Output

```
# Press Clipping Report - January 17, 2026

**Generated**: 06:45 AM UTC
**Search Terms**: Media XYZ, Product ABC, CEO John Doe
**Timeframe**: Last 24 hours (primary) / Last 72 hours (context)
**Sources Scanned**: 47 / 50 successful

## Executive Summary
- **Total Items Found**: 23 articles, 45 social media posts
- **Trending**: Product ABC launch coverage (18 mentions)
- **Alert**: Negative sentiment detected in 3 major outlets
```

Document Version: 1.0
Last Updated: January 17, 2026

Owner: [Your Team/Organization]
Status: Draft for Review