Web-Intelligence WS 25/26

Name: Irmak Damla Özdemir **Matrikel-Nr.:** 5123127

Illustrieren Sie das Problem von Big Data an Ihrem Anwendungsfall. Wenden Sie eine Lösung des Problems für Ihren Anwendungsfall an.

Trending Health News: Problem of Big Data

In my project, the system collects health-related news articles from several online sources such as Spiegel Gesundheit, Tagesschau Gesundheit, WHO, and Deutsches Ärzteblatt.

Each article includes metadata such as title, summary, publication date, link, and source. Because these sources publish new articles frequently, the system quickly generates large, fast-growing, and diverse datasets. This leads to a classical Big Data problem.

Large Amounts of Data

Every source publishes multiple articles daily. Over weeks and months, this results in thousands of entries, each containing: long text (like summaries), many metadata fields, multi-language content (German and English from WHO) and multiple CSV exports with timestamps.
As the dataset grows, storing, cleaning, and analyzing it becomes more complex.

Heterogeneous Data

The sources differ in:

- Topic scope
- Writing style and terminology
- Language (German vs English)
- Publication patterns (frequent vs scheduled)

This variety makes topic detection harder, since keywords must match both German and English expressions e.g., “depression”, “depressiv”, “anxiety”, “Angst”.

Continuous Data Flow

RSS feeds deliver new articles in real time, especially Tagesschau, which updates the most frequently among the chosen sources.

When it is automated, the system must:

- Fetch new data regularly
- Validate missing or inconsistent fields
- Detect topics automatically
- Store results for visualization

This creates processing pressure typical for Big Data systems.

Concrete Big Data Challenges in the Project

Keyword detection becomes inaccurate due to language differences, especially because WHO publishes in English while other sources use German keywords, and also because the current keyword lists are not comprehensive enough. As a result, many relevant terms cannot be matched correctly, which leads to an accumulation in the *Other* category during topic detection.

Uneven publishing frequency skews trend results.

Tagesschau dominates the dataset with the most frequent updates, making trends appear biased unless normalized.

Applied Big Data Solution for My Use Case

To handle the Big Data challenges in my use case, I apply a lightweight approach based on sampling and aggregation. Instead of storing every single article from all sources, the system focuses on collecting only a daily sample of the most relevant headlines from each provider. This reduces data volume while still preserving the overall trend structure.

Additionally, basic aggregation techniques are used: articles are grouped by publication day and by detected keyword category. This makes the dataset much smaller and easier to analyze without requiring complex Big Data infrastructure.

To reduce the impact of language differences, a simple language-normalization step is added. Before keyword matching, titles and summaries can be translated into a single target language. This prevents mismatches between German and English expressions and reduces the number of articles classified as *Other*, making trend detection more accurate.

This approach solves the Big Data problem by minimizing the amount of data while keeping the ability to identify meaningful health trends.