**Deliverable 02**

**Project Modules**

1. Data Acquisition Source

* DAWN NEWS (ePaper)

1. Preprocessing of Unstructured Data

* Case Folding
* Stop Word Elimination
* Lemmatization

1. News Classification
2. Information Extraction

**Different Methods Considered for Data Acquisition**

* Google Reader
* Service unavailability from 1st July
* Only few days/hours previous data can be gathered through
* Mozilla Thunder Bird
* Dependency on third party
* Difficulty in parsing dawn source file

**Data Acquisition Methodology**

* The method that is followed for the data acquisition is parsing the provided links automatically.
* Downloading the inner html through required fields.
* **HTML AGILITY** pack is used for parsing the html and getting the required inner html.
* This unstructured text is then, preprocessed. Preprocessing includes;

1. Case Folding
2. Eliminating Stop Words
3. Lemmatization