

AML-2304

Natural Language Processing

Assignment

(20% of Final grade)

|  |  |
| --- | --- |
| **Instructor:** | Vahid Hadavi |
| **Class:** | **AML-2304** |

1. **Dataset Collection**

You have the option to choose data either about financial news or amazon product reviews or other sources that have text data. Some of the available resources you can refer to are the Python libraries in the following: (You have the choice to research and implement from other options)

* Fin-news: <https://pypi.org/project/fin-news/>
* Sample of the Amazon Customer Reviews Dataset provided by tensorflow: <https://www.tensorflow.org/datasets/catalog/amazon_us_reviews>

1. **Important Notes**

* The group report is to be submitted in a Jupyter Notebook with Python code and explanation paragraph explaining your thought process in the PDF report and also provide the data used for this project

1. **Task Details and Evaluation Criteria**

Explain your data sources. Create a Regressor or Classifier depending on choices using text data. Try different vectorization methods for your solution and explain the comparison between 2 different word vectorization methods you choose (e.G tfidf, bag of words etc)

Your choice and interpretation of the word vectorization methods are up to you but please state your rationale and assumptions. The submission will be considered with regards to

• methods,

• analysis that should cover the aspects of

* + - Dataset collection explanation
    - word-clouds that are built based on for each method of vectorization
    - Dataset preview and summary statistics introduction and EDA
    - Data cleaning and text preprocessing explanation
    - Word vectorization methods
    - Model building process and results interpretation and comparison (min 2 models needed)

• programming,

• communication skills,

• quality of results,

• and discussion of results.

# Your submission should include proper visualization and use the template provided on the Moodle.

# Only one person per team should submit the deliverables.

1. **Deliverables:**

The deliverables for the final result should be:

1. A Jupyter Notebook including the code chucks and explanation paragraph step by step
2. PDF report in standard technical format as discussed in the class
3. data