2024 :: Advanced Topics in Data Science / Data Mining II :: Practical Assignment

Analysis of Posting Strategies for Higher Education Institutions

Final project submission deadline: May 24th @ 23:30 (on MoodleUP)
Project proposal presentation: May 27th, during the lectures.

Important:

- Be aware of the University of Porto Policy on Plagiarism.
- Properly cite any external source that you used.
- You are **not allowed to** use or reuse any piece of code **created by the other groups**.
- However, you can use any programming language for the development of your work, as well as code that is freely available (open source) on the Internet (provided you understand it, and properly refer to it).

General description:

You are provided with a subset of a dataset comprised of the X/Twitter posts from Higher Education Institutions (HEIs) for the purpose of conducting specific analyses and making predictions. It is important to note that you must work with this given dataset and not obtain it from any other source.

The objectives of this project are twofold:

i) First, you are expected to conduct a comprehensive exploration of the data to gain insights into the patterns of each HEI, for example the frequency and periods during which they typically occur (among other features). Your findings should be presented using visualizations. You are then required to make groups of HEIs that share commonalities in their posting patterns.

ii) In the second part of the project, you are required to identify the main topic for each post. The objective is to categorize the posts into following the categories {“Image”, “Education”, “Research”, “Society”, “Engagement”}. It is up to the group to decide how to do this. Then, train and use at least two machine learning models, to predict the category of the next post. You can also try association rules for this purpose. These models should then be evaluated against actual data to assess their performance.

In detail:

Part I

- Analysis of post frequency, time of day, and day of the week.
- Size and type of the posts.
- Group HEIs according to common posting patterns.
- Identify the differentiating features.

Part IIa

- Analyze post text content.
  - the sentiment of the posts.
  - the emotions.
  - the important words.

Part IIb

- Identify the main category for each post.
- Create a system to predict the next post’s category after 20+ posts.

Remaining activities:

- Report your process in an article-type document (IEEE template, **max. 5 pages**)
- Do a presentation of your work.

Deliverables:

1. All the material needed to perform the analysis, prediction and the created plots/figures (R code)
   Filename: `Group_#_code.zip`
2. The project report in the form of an article in PDF
   Filename: `Group_#_article.pdf`
3. An electronic presentation (preferably in PPT/PPTX)
   Filename: `Group_#_presentation.(pptx)`

1. **[20 points] The dataset exploration and creation of plots (Part I)**
   a. Perform exploratory data analysis and do all necessary data cleaning.
   b. Use visualizations to better and more rapidly understand the data.
   c. Cluster the HEI according to your chosen parameters and analyze the results.

2. **[30 points] Content Analysis (Part II)**
   a. Perform sentiment analysis and emotion recognition on the posts of each HEI.
   b. Use the TF-IDF methodology to identify important words.
   c. Devise a way to use those words to categorize the posts in the provided categories.
   d. Create a ML model to predict the next category. You are expected to create a model with at least 8 features. Also, for each feature, there must be a particular motivation for including it in the model (which you should briefly describe in your report).
   e. You should also try to categorize using association rules.
   f. Evaluate the selected algorithms according to well-known metrics.

   Be advised: it is important that your group can add something that differentiates itself from any other group in the class (e.g., a different model, different features). You will score for that creativity.

3. **[15 points] Report as an article**
   a. The project report should be written in the standard IEEE conference format template (**max. 5 pages**) and submitted in PDF.
   b. The report should include the title, authors, an abstract and a conclusion (at least). The inclusion in the written report of at least 4 figures is mandatory.
   c. Explain the transformations you made to the dataset in order to obtain your final model, discuss the features you use to tackle the problem and your novelties.
   d. Explain how you categorize a topic.
   e. Report on the evaluation of your model.

   Note: reports that do not strictly follow the template will not be considered for assessment.

4. **[15 points] Final presentation**
   a. Prepare a 15-minute (maximum) presentation to explain what you did and the results you achieved using an electronic presentation as a support, mainly for graphics, illustrations, and tables. Note that every group member must participate in the presentation.
   b. You can schedule your presentation for one of the slots on an activity in MoodleUP (to be available). Please, choose a single group participant to register the whole group choice, as the number of time slots is limited.

   Note: You will get points based on the **visual quality** of the electronic presentation.