Data mining project

Information Operations

A.Y. 2026/2027

A **project** consists of data analysis based on data mining tools. The project has to be performed by a team of 3 students. It has to be performed by using Python. The guidelines require addressing specific tasks, and results must be reported in a unique paper. This paper's total length must be around **25 pages** (precise length of the report will be communicated according to the selected tasks) of text including figures. The students must deliver both: paper and well-commented Python notebooks. Here you can download the dataset and the latex report template for the report: do not modify the font, font size, and margins.

Dataset ZIP: Download Here

Latex Report Template: Download Here

Dataset description

Information operations are coordinated attempts to manipulate or influence public discussion online. The source data include posts from accounts associated with verified information operations and posts from control accounts discussing related topics in comparable time windows.

The goal is to study how coordinated accounts differ from control accounts, and how these differences can be represented through features.
The dataset contains the original post and account fields, of several Information Operations, carried on by different countries.

In the dataset there are the following variables:

1. **ioCountry**: Information operation country.
2. **postid**: Anonymized unique identifier for the post.
3. **post_text**: Textual content of the post.
4. **application_name**: Application or client used to publish the post.
5. **post_language**: Language label associated with the post.
6. **in_reply_to_postid**: Identifier of the post replied to, when the post is a reply.
7. **in_reply_to_accountid**: Identifier of the account replied to, when available.
8. **post_time**: Timestamp of publication.
9. **accountid**: Anonymized identifier of the account that published the post.
10. **account_profile_description**: Profile description of the account, when available.
11. **follower_count**: Number of followers recorded for the account.
12. **following_count**: Number of accounts followed by the account.

13. **account_creation_date**: Creation date of the account.
14. **is_repost**: Boolean field indicating whether the post is a repost.
15. **reposted_accountid**: Identifier of the reposted account, when applicable.
16. **reposted_postid**: Identifier of the reposted post, when applicable.
17. **hashtags**: List of hashtags appearing in the post.
18. **urls**: List of URLs appearing in the post.
19. **account_mentions**: List of account mentions appearing in the post.
20. **is_control**: Class label. True indicates a control post; False indicates an information operation post.

Task 1: Data Understanding and Preparation (30 points)

Explore the Information Operations dataset using the analytical tools studied during the course and write a concise data understanding and preparation report.

The analysis should include:

- **Data semantics**, describing the meaning of the variables considered in the analysis.
- **Data quality assessment**, considering missing values, errors and inconsistencies, duplicated records, and outliers.
- **Distribution of the variables and descriptive statistics**, supported by appropriate visualizations.
- **Pairwise relationships and correlations**, with attention to possible redundant information.
- **Feature engineering**, defining meaningful indicators that capture relevant characteristics of the phenomena under analysis.

User Profiling and Feature Engineering

Starting from the post-level dataset, create a profile **for each user** based on their posting behavior. Improve the quality of the data and derive new features that are informative for describing the users.

Examples of possible indicators include:

- Compare each user's number of posts with the typical activity of users in the same `ioCountry`, to identify unusually active accounts.
- Measure how concentrated a user's activity is over time, for example through the share of posts published during their most active day or week.
- Compute the entropy of the distribution of `post_language` for each user, to characterize whether they tend to use one or multiple languages.
- Compute the entropy of the distribution of `application_name` for each user, to characterize whether they rely on one or multiple posting clients.
- Combine information about replies, mentions, and reposts to characterize different patterns of interaction and content broadcasting.

These examples are **not mandatory**. Each team can define different indicators that it considers meaningful for characterizing users and their posting behavior.

It is **mandatory** that each team defines a set of derived indicators. Each indicator must be accompanied by a clear description of what it is intended to capture and, when appropriate, by its mathematical formulation.

The resulting user profiles and derived features will be used for the clustering analysis in **Task 2**.

The data understanding process should accompany the different stages of data preparation: the relevant analyses should therefore be carried out not only on the original data, but also on the resulting user profiles and derived features.

Task 2: Clustering analysis (30 + 2 optional points)

Based on the user profiles and features obtained in Task 1, explore the users of the Information Operations dataset using different clustering approaches.

Select an **appropriate set of attributes** for the clustering analysis and motivate your choices.

For each clustering approach, investigate the relevant parameters and configurations and **characterize and interpret the obtained clusters**. Algorithm-specific information, such as cluster centroids, should be considered when appropriate.

The following clustering approaches must be considered:

- **K-means clustering**
  - investigate the choice of the number of clusters (k);
  - characterize and interpret the resulting clusters.
- **Density-based clustering**
  - investigate the relevant parameters of the selected algorithm;
  - characterize and interpret the resulting clusters, including the role of noise when relevant.
- **Hierarchical clustering**
  - investigate and compare different configurations of the algorithm;
  - visualize and discuss appropriate dendrograms;
  - characterize and interpret the resulting clusters.

Finally, **compare and evaluate the clustering solutions obtained with the different approaches**, discussing their similarities and differences and the strengths and limitations of the different methods. Quantitative clustering evaluation measures may support the analysis but should not constitute the sole basis for the comparison.

Optional task (+2 points)

Explore an additional clustering technique and discuss whether it provides additional insights with respect to the mandatory clustering approaches. You can use this package or another of your choice: https://github.com/annoviko/pyclustering/

Task 3: Time Series and Multi-view Analysis (30+2 points)

For this task, select a single **Information Operation (IO)** country of your choice and motivate your selection. All subsequent analyses should focus on the users belonging to the selected IO.

Task 3.1: Time-Series Construction

Create a dataset of **univariate time series** representing the posting behavior of each user.

Define an appropriate indicator whose values over time will constitute the time series. Clearly describe what the indicator is intended to capture, how it is computed, and how its values should be interpreted.

Motivate the methodological choices adopted for constructing the time series, including the temporal granularity, observation period, and any preprocessing or transformations applied.

Task 3.2: Time-Series and Multi-view Clustering

Perform clustering on the users belonging to the selected Information Operation, exploiting their **temporal and tabular representations**.

The analysis must include:

1. **Time-series clustering:** apply an appropriate clustering algorithm to the univariate time series constructed in Task 3.1.
2. **Tabular clustering:** apply at least one clustering algorithm to the tabular user profiles, considering only the users belonging to the selected IO. This analysis is distinct from Task 2, where clustering was performed on the entire dataset.
3. **Multi-view clustering:** combine the temporal and tabular representations using at least one approach selected among **early fusion, late fusion, and joint multi-view clustering**. Consider possible redundancy and complementarity between the two representations.

For each approach, characterize and interpret the obtained clusters, investigating the differences in the distributions of the relevant variables across clusters and with respect to the overall population.

Finally, **compare, characterize, and discuss the clustering solutions** obtained using the temporal, tabular, and combined representations, highlighting their similarities, and differences.

Optional task (+2 points)

Apply a **second fusion approach**, different from the one selected in point 3, and compare the resulting clustering solutions, discussing whether the additional approach provides different or complementary insights.

Task 3.3: Topic Modeling

Apply **BERTopic** to investigate the topics discussed in the posts of the selected Information Operation.

The analysis must include:

1. **Topic modeling on the entire selected IO:** apply BERTopic to the complete collection of posts belonging to the selected IO. Identify, characterize, and interpret the main topics discovered.
2. **Topic modeling within clusters:** select one of the clustering solutions obtained in Task 3.2 and apply BERTopic separately to the posts published by the users belonging to each cluster. Consider only clusters containing a sufficiently large number of users and posts to support a meaningful analysis, defining and motivating appropriate minimum thresholds.
3. **Comparison and characterization:** compare the topics discovered within individual clusters with those obtained from the entire IO, discussing similarities and differences in their thematic content. Further characterize the discovered topic groups by investigating their relationships with relevant user characteristics, such as the **is_control** label.

Discuss whether the cluster-specific topic analyses reveal differences in the content published by users with different behavioral profiles.

Rules for final delivery and Exam

Project Delivery. The final deadline of the project is **4th January 2026 at 23:59**. This deadline is **STRICT**. Groups that will not deliver the project by 4th January will need to do the written exam.
Each group must upload the project (code and report) by using this

Google form: https://forms.gle/XMbZNAGRbQdA8zQ98

The uploaded project must be a zipped folder named **DM_GroupID.zip** and containing 3 folders and 1 pdf file:

1. a folder named **DM_GroupID_TASK1**
2. a folder named **DM_GroupID_TASK2**

3. a folder named **DM_GroupID_TASK3**
4. a pdf file with maximum **25 pages** including figures discussing the results of the tasks. The name of this file must be: **DM_Report_GroupID.pdf.** The file must contain the list of authors (i.e., members of the group).

How to book for the exam colloquium?

In https://esami.unipi.it/ you can find the dates for the exam: one for January and one for February. Each student must do the registration on one of the 2 dates. These are not the dates of the colloquium but we will use the list of registered students for organizing the exam dates. We will share with you a calendar for the oral exam.