Text Mining and Analytics, Spring 2024, Course Project

Instructor: Behrooz Mansouri (behrooz.mansouri@maine.edu)

Text Mining and Analytics course at the University of Southern Maine is a project-based course. Students will work, in groups of 2 or more, on a project of their choice. The project is broken into three sections, described later in the document.

Learning Objectives: The goal of this project is for students to do a text mining research and contribute to the field. Throughout the course, they will learn about different techniques for mining text documents, evaluating and analyzing the results. In the project, it is expected that students will apply those techniques to their problem.

Project Topics: Students can choose the topic of their interest. However, it is strongly recommended to explore labs at CLEF 2024. Upon completion of the project, students can submit their papers to CLEF labs and participate in the conference with the instructor's support. For other topics, students should get the instructor's approval. The criteria for approval is having an existing dataset for evaluation or a clear path to create one and now evaluation measure (i.e., who would you measure the success of your project). Other criteria for acceptance include the following:

- Relevance: The project should be relevant to the course objectives and align with the current state-of-the-art in text mining research.
- Originality: The project should be innovative and contribute to the existing text mining research in some way.
- Feasibility: The project should be feasible within the given time frame and resources available, and the project plan should demonstrate that the students have a clear understanding of the steps involved in completing the project.
- Clarity of objectives: The project goals and objectives should be clearly stated and well-defined so that they can be evaluated effectively.
- Technical proficiency: The students should demonstrate that they have the technical skills necessary to complete the project, or a plan to acquire these skills if necessary.

While the projects are done in teams, each student is responsible for one specific task in the project. All the codes should be submitted on Git platforms (GitHub or GitLab). All the reports will be delivered written in Overleaf (Word documents are not acceptable). Note that students are not allowed to change the topic/team after the proposal is accepted.

Here is a description of each project section:

Part I: **(The week before Spring break)**

In Part I, students will only explore existing approaches for the task they are tackling. The requirement is for students to read a minimum of 8 papers and have a full understanding of the task and existing models to solve the problem. It is recommended that students include 1–2 papers that provide an overview/survey on the problem to first fully understand the problem. Also, at least four papers should be published in the past two years.

This phase is very important for the project as it gives the students an idea about what the problem is, and also provides insight into the current existing approaches to tackle the problem. Upon completing this part of the project, students will have a better understanding of the next steps. For deliverables, students will write a minimum of three pages of their understanding of the paper and the approaches. In this document, they should discuss the strengths and weaknesses of each approach they read, providing a table. In addition to the report, students will have class presentations to introduce approaches to their classmates.

It is essential to choose the right papers for this phase; Papers that are peer-reviewed, have publicly available code (their results are reproducible), and have the right experimental settings. During the course, students will learn how to find papers.

Part II: **(First week of April)**

In this part, students will explore one of the existing systems for their project. The goal is to provide a baseline system for Part III. In the rare case that there are no existing systems for the problem, students can use a baseline model of their choice. For example, if you decide to work on legal information retrieval, and there are no systems available, you can use the TF-IDF model as your baseline which is a general information retrieval system, not specific to a domain.

The objective is to run the baseline system on the project data, get the evaluation results, and do a proper analysis of the results. Students should be able to discuss what worked and what did not, by providing examples for each. Then they should provide an analysis of what they think can be improved and form a research question/hypothesis for Part III of the project.

The deliverables for this phase include a minimum of three pages of report, codes of Git, and a

short presentation to the class. This phase also includes in-person delivery to the instructor during the office hours.

Part III: **(Final exam time)**

In the final part of the project, students will implement their solution for the problem. This phase aims to answer the research question formed in part two. It is important to note that, you are not expected to have state-of-the-art results, or evaluation measures higher than existing models. You will be assessed on how well you have explored your results and why your research hypothesis is rejected or accepted. The deliverables of this phase are similar to those of Part II except for the report.

For report delivery, students should work with their team and put all the reports (from different phases) and different tasks into one paper, with a minimum of 12 pages.

After Course:

During the course, students will complete research on text mining, they have the opportunity to work with the instructor and publish their research. This is a great addition to your CV. This part is not mandatory, and it is only for interested students.

Notes:

- For each phase of the project, students will receive instructions on deliverables and expectations
- For any questions, please use the project channel on the course Slack group
- Students are highly encouraged to use all the available resources online, including Chat-GPT, BARD, ChatPDF, and others. If you have found useful tools, please share them with the class on the Slack channel
- While teamwork is encouraged in the class, each student is responsible for their task, and the assessment is done individually. No student is responsible for their team members' task
- For the reports, we will use Springer's LNCS template