University of Windsor

School of Computer Science
Faculty of Science

COMP-8730: Natural Language Processing & Understanding
Winter 2021

#	Title	Due Date	Grade Release Date
1	**Proposal**	Jan. 25, 2021, AoE	Feb. 01, 2021, AoE

This course is research-oriented and project-driven in which a research project should be defined and completed in the field of NLP within one semester. The objectives of the research project are to provide graduate students with:

- An experience with research procedure, in general, and research in NLP, in particular.
- Hands-on experience with NLP.
- Advancing state of the art in NLP while passing a grad course.
- An opportunity to present a research outcome in an international computer science conference
- An opportunity to meet with scholars in the NLP community

In the research project, we propose a solution(s) to a problem by implementing an algorithm like a software project. However, there are differences in some respects. For instance, while a software project may implement an existing algorithm, a research project should propose and implement a *new* algorithm that improves or addresses a particular aspect of a problem that the current algorithms overlook. Roughly, a research project has the following milestones (phases):

1) Proposal
2) Literature Review
3) Proposed Method (Formal + Code)
4) Experiment (Evaluation)
5) Presentation (Paper + Talk)

In this course, a manual is prepared to guide the students through each milestone. The current manual is for the first milestone: Proposal, which further has the following steps:

Introduction

The 1st step of the Proposal is to select a research problem to advance science in NLP. Usually, a research problem is already defined, and solutions have been proposed to address the problem. Research problems are always open for new solutions that either have better answers (results), run faster (time complexity), or need a simpler running platform. For instance, look at a survey of all available solutions in sentiment analysis of tweets on Twitter by Giachanou et al.

You might come up with a problem that you think is original, and no one has faced the problem before. Although you might be right at first, a further search would prove that it is not true since most of the problems are either the same while in a different form or related in one way or another. For instance, you may come up with the idea of identifying the author's gender for tweets on Twitter. You may think that no one has solved this problem and you are the first to propose a solution for it. However, if you search, you will figure out that this problem is an instance of document classification, and researchers have already proposed solutions. This does not mean that you should quit! If you read their solution, you would see that the solution is for identifying the gender of the book's authors, which is marginally different from what you want to solve. So, you can continue on this problem, arguing that if those methods are applied on tweets, they will yield poor results since they assumed that the documents are books that are

inherently different from tweets. Books are long formal documents, while tweets are very short informal pieces of text and need special care. Further, there are research problems that no solution has been proposed, or all the proposed solutions were not successful. For this course, please do not choose them as a research project.

In this course, you can find a problem from the following list or ask for the instructor's approval of your own choice of the problem otherwise. You are not allowed to share the same problem with other students unless you are a team. You will need to communicate your choice with the instructor and obtain approval.

Area	Sub Area
Social Media Mining	User Interest Modeling User Community Detection Event Detection Friend Recommendation
Natural Language Misunderstanding (De-biasing)	Gender Bias Racial Bias Other types of Bias
Fake Information Detection	Fake News Detection
Abusive Language Detection	Hate Speech Detection
Web Search	Social Query Expansion
Information Extraction	Explicit Entity Detection Implicit Entity Detection Semantic Annotation
Multimodalities (Visual + Audio)	
Sentiment Analysis	
Dialogue and Interactive Systems	
Cognitive Modeling and Psycholinguistics	
Translation and Multilinguality	
Question Answering	
Summarization	
Syntax: Tagging, Chunking and Parsing	

Motivation

The 2nd step of the Proposal is to find motivations for solving the chosen research problem. In this course, we do not do a research project for the sake of science per se but for the benefit of human society. For instance, if you choose to find the gender of the author of a tweet as a research problem, you have to ask, "So what?", "Why should others and I put our effort into solving this?", "What are the benefits for human society?" You would say that knowing the gender of an author for a tweet would help with 1) better friend recommendation as shown in [*, *], 2) better product recommendation as shown in [*], 3) health issue prediction as shown in [*]. The * are citations to the scientific papers!

Then, you provide a reasonable motivating example that clearly explains the benefit. This part might be easy since all the other existing solutions mention these in some way. For instance, you would say:

> *For instance, given the tweet "Don't know what rain jacket to buy" by @prince" if we could find the gender of the author, e.g., male, we are able to recommend rain jackets that are made for men.*

Also, you have to motivate why you want to continue solving the problem. Do you want to improve the solutions? How? Why? You would say:

> *Although there are methods that identify the gender of an author for a book [*] or a paper [*], there has been no work, to the best of our knowledge, that does so on tweets. Tweets are short, noisy, and informal, different from formal documents such as books and papers.*

Or if there is an approach that does so on tweets, after a careful reading of that work, you understand that the method falls short in some exceptional cases, you would say that:

> *Although there are methods that identify the gender of an author for a book [*] or a paper [*], there have been few works that do so on tweets like [A]. However, the method proposed in A uses the author's name to find the gender, which falls short when the author's name is unisex like 'Andy' or is 'Empr_World,' which does not have any gender implication.*

Problem Definition

The 3rd step of the Proposal is to formalize your problem with a language that is clear without any ambiguity to all researchers in computer science. In this course, this language is math. For instance, in the problem of finding the gender of an author for a tweet, we should formalize it in math like:

> ***Author Gender Identification per Tweet:*** *given a corpus M and a tweet $t \in M$ that is posted by a user u, our task is to find u's gender from the set {-1:male, 0:na, +1:female}*

Your formal problem definition should be followed by an intuitive example like:

> *For instance, given the tweet "Hey Georgia! If you still have your mail-in ballot, return it to a drop-off location before 7 PM ET today. Find one near you at http://weall.vote/ga. Let's get it done! Police cars revolving light" by @MichelleObama," we want to output +1:female.*

Author list

The 4th purpose of Proposal is to create a team of maximum 2 students to research experience collaboration and conflict resolution. Although you can do a research project individually and alone like Dr. Youcef Derbal, collaboration is highly encouraged since it helps with better workload distribution and sooner and better fault detection, to name a few. In a team research project, the contribution of the members for each part of the research project must be clear. Evaluation scores will be distributed among the team members based on the significance of the contributions.

In summary, the Proposal has (%marking schema for this milestone):

1.1) (%20) Introduction
1.2) (%30) Motivation to solve the problem + an example
1.3) (%40) Problem formal definition + an example
1.4) (%10) Team members (author list including affiliation)

Submission Guidelines

- Submission must be written in English, in the current ACM two-column conference format in LaTeX. Overleaf templates are available from the ACM Website (use the "sigconf" proceedings template).

- Submission must be 1 page in length, no more not less, including figures, tables, references, authored by the team members.

- Submission must be in one single zip file with COMP8730_Proposal_UWinId1_UWindId2.zip, including:
  1. the LaTeX files
  2. the pdf file

A sample submission has been attached to this manual in Blackboard, also available online here.