CS 579: Online Social Network Analysis

Project II

Prof. Kai Shu
Due on November 30, 2020 at 11:59 pm

This is a ***2-member*** project assignment. Each group is supposed to work on the steps together, including the writeup of the report. Please submit one set of your report and related files per group on Blackboard.

We suggest you to form a different team from Project I so that you can make more friends :)

1 The Default Project

All undergraduate students should work on this project. As for Ph.D. and Masters students, there are two other options available:

1. Proposing your own research projects.
2. Working with Prof. Shu and TAs for promising research projects.

You still need to form a ***2-member*** group for the above options. You can find the instructions in Section 2 and Section 3.

1.1 Task: Fake News Classification

Social media has become one of the major resources for people to obtain news and information. For example, it is found that social media now outperforms television as the major news source. However, because it is cheap to provide news online and much faster and easier to disseminate through social media, large volumes of fake news or misinformation are produced online for a variety of purposes, such as financial and political gain. The extensive spread of fake news/misinformation can have a serious negative impact on individuals and society: (i) breaking the authenticity balance of the news ecosystem; (ii) intentionally persuading consumers to accept biased or false beliefs; and (iii) changing the way people interpret and respond to real news and information. Therefore, it is important to detect fake news and misinformation in social media.

We formally define the task as follow. Given the title of a fake news article A and the title of a coming news article B, participants are asked to classify B into one of the three categories:

- **agreed:** B talks about the same fake news as A.
- **disagreed:** B refutes the fake news in A.
- **unrelated:** B is unrelated to A.

1.2 File Descriptions

In the attached folder, you are provided with 3 CSV files:

- **train.csv:** Training data
- **test.csv:** Test data
- **sample_submission.csv:** Expected submission format

The training data includes the "label" of each news pair, while the test data doesn't. Validation data can be split from **train.csv**. Students should use the training data to train a classifier and evaluate their model's performance with the validation data. Finally, by using the trained model, you are required to predict the results for the test data. The format of your output file should be the same as "sample_submission.csv" with your prediction replaced in"`lable`" column.
The columns in train and test data are as follows:

- **id**: the id of each news pair.
- **tid1**: the id of fake news title 1.
- **tid2**: the id of news title 2.
- **title1_en**: the fake news title 1 in English.
- **title2_en**: the news title 2 in English.
- **label**: indicates the relation between the news pair: agreed/disagreed/unrelated.

1.3 Submission

Students are supposed to submit the result file (named ***"submission.csv"***), source code, presentation slides and report in one *.zip* file named LASTNAME1_LASTNAME2_PJ2 (Instead of LASTNAME1 and LASTNAME2 type the lastname of each member).
The submitted results should be reproducible with the submitted code/data. Moreover, do not change the name of the files as your submitted .csv file will pass an automatic program.
The report should not be less than 2 pages and should include description of the data pre-processing, model, and validation results.
Use a "Reference" section and cite all the papers, tutorials, packages, software and libraries you used for your program.
The class on Nov 30 would be for students' presentations.

2 Proposing Projects by Your Own

For MS students with thesis and PhD students, if you would like to work on projects related to your thesis, you can propose your own project statement. Please submit your proposal in Blackboard before Oct 11 (Proposal template would be provided). Please be careful when submitting proposal. Do not submit it under Project II, there would be an individual window for Project II proposal. We will review the proposal statement for approval.

3 Working on Cutting-Edge Research Projects

For MS students with thesis and PhD students, if you want to work closely with Prof. Shu for promising research projects, you can choose from following projects with the corresponding contact information:

- **Learning with weak social supervision**: Limited labeled data is becoming the largest bottleneck for supervised learning systems. This is especially the case for many real-world tasks where large scale annotated examples are either too expensive to acquire or unavailable due to privacy or data access constraints. Weak supervision has shown to be a good means to mitigate the scarcity of annotated data by leveraging weak labels or injecting constraints from heuristic rules and/or external knowledge sources. We will explore how to use weal social supervision for various prediction tasks. Contact **Hao Ding** at hding9@hawk.iit.edu or **Lan Wei** at lwei3@hawk.iit.edu

- **Human-in-the-loop disinformation detection**: Consuming news from social media is becoming increasingly popular. However, social media also enables the wide dissemination of disinformation including fake news. Because of the detrimental effects of fake news, fake news detection has attracted increasing attention. The fake news can take advantage of multimedia content to mislead readers and get dissemination, which can cause negative effects or even manipulate the public events. One of the unique challenges for fake news detection on social media is how to identify fake news on newly emerged events. Unfortunately, most of the existing approaches can hardly handle this challenge, since they tend to learn event-specific features that can not be transferred to unseen events. We will explore how to leverage domain knowledge from knowledge bases (e.g.,Wikipedia) to help the fake news prediction with little labeled data. Contact **Hao Ding** at hding9@hawk.iit.edu

- **Few-Shot learning for disinformation detection.** Recent studies have usually used deep neural network models whose performances depend on the amount and quality of training data. For fake news detection, there are few labeled data and a large amount of unlabeled data in real scenarios. One prominent way is to consider few-shot learning from the perspective of data. Therefore, we would like to investigate how to exploit few-shot learning to detect fake news more effectively with limited labeled data. Contact **Lan Wei** at lwei3@hawk.iit.edu

- **Fairness in recommendation systems.** Modern collaborative filtering algorithms seek to provide personalized product recommendations by uncovering patterns in consumer product interactions. However, biased data can lead collaborative-filtering methods to make unfair predictions for users from minority groups. Therefore, we would like to study how to perform fairness-aware recommendations effectively with limited sensitive user attributes. **Zhenghao Zhao** at zzhao48@hawk.iit.edu