CSE 635: NLP and Text Mining

Spring 2020
Instructor: Rohini K. Srihari

Class Project Description and Requirements

Overview

The goal of this semester long project is to provide hands on experience designing, implementing, evaluating and demonstrating a complete web mining/text mining/social media mining solution based on a combination of natural language processing (NLP), information retrieval (IR) and machine learning (ML) techniques. You are provided a choice of three topics which broadly fall into the area known as AI for Social Impact. The three topics cover fake news, social disruption, and social media mining in the healthcare domain. Each of the projects will have a standard dataset and ground truth enabling quantitative evaluation. Many of these are from past or ongoing challenges and have been attempted by other teams. We encourage you to use any available online tools or platforms to develop your solution. You should strive to produce results that would be in the top 10% of any previously published results on the same dataset.

While there is a quantitative evaluation component on a static data set, we are also requiring you to develop a live system using current/live data. This will also involve developing a visual user interface so you can demonstrate the system.

This project will satisfy the MS project requirements specified by the CSE department. While the problem definition and evaluation dataset have been fixed, there is ample room for creativity on your part in further enhancement of the solution, and implementation. Be creative, and most importantly pace yourself properly during the semester.

Your project is divided into three phases which are described in more detail later on in this document:

Phase 1: Submission of project proposal and in-person presentation of your proposal. This includes a comprehensive literature review on your selected topic, a necessary step before you begin the design of your own system!
Phase 2: Interim report describing evaluation on baseline system.
Phase 3: Final submission of technical paper and in-class presentation of your end-to-end system.

Project Option 1: Fact Extraction and Automated Claim Verification

Due to the deluge of information on the internet, there has been an exponential increase in online misinformation and rumor mongering. For this reason, there is a dire need for automated verification of claims or fact-checking. In this project, students are required to verify textual claims against textual sources by retrieving relevant evidence from Wikipedia pages. Students are required to use a benchmark dataset FEVER (Fact Extraction and VERification) which consists of ~185k claims manually verified and annotated by expert annotators. The dataset and the tasks involved to complete this project are detailed below:

Dataset: FEVER dataset which contains 185,445 claims manually verified against the introductory section of Wikipedia pages and classified into SUPPORTED, REFUTED and NOTENOUGHINFO. For the first two classes, the claims are associated with necessary evidence supporting or refuting the claim. Hence, the training data contains 4 fields:

- id: The ID of the claim.
- label: One of the {SUPPORTED ; REFUTED ; NOTENOUGHINFO}
- claim: The text of the claim
- evidence: A list of evidence sets (lists of [Annotation ID, Evidence ID, Wikipedia URL, sentence ID] tuples) or a [Annotation ID, Evidence ID, null, null] tuple if the label is NOT ENOUGH INFO. (the Annotation ID and Evidence ID fields are for internal use only and are not used for scoring.)

a. **Training Dataset:** https://s3-eu-west-1.amazonaws.com/fever.public/train.jsonl
b. **Development Dataset (Labelled):** https://s3-eu-west-1.amazonaws.com/fever.public/shared_task_dev.jsonl
c. **Development Dataset (Unlabelled):** https://s3-eu-west-1.amazonaws.com/fever.public/shared_task_dev_public.jsonl
d. **Test Dataset:** https://s3-eu-west-1.amazonaws.com/fever.public/shared_task_test.jsonl

You can find more information and examples on the dataset by visiting this website: http://fever.ai/2018/task.html

Task Definition: You are required to build an end-to-end system by completing the following tasks:

- Find and retrieve Wikipedia pages which are *most relevant to the claim.*
- Extract a set of sentences from the retrieved Wikipedia pages that support or refute the claims. These set of sentences form the evidence for the claim.
- Using this evidence, classify the claim as **Supported** and **Refuted**.
- If there isn’t sufficient evidence to support or refute a particular claim, label the claim as **NotEnoughInfo**.
- Please note that a claim’s evidence may contain *multiple sentences* which are to be examined together to provide correct label. For example, a claim “Oliver Reed was a film actor” one piece of the evidence can be a set of two sentences: {“Oliver Reed starred in the Gladiator”, “Gladiator is the film released in 2000.”}

Evaluation Metrics: To evaluate your system, you are required to use the FEVER scoring program which can be found at this GIT repo: https://github.com/sheffieldnlp/fever-scorer. We are primarily interested in the *accuracy* metric for the classification task and *precision, recall and F1* metrics for evidence extraction. You are encouraged to understand the FEVER scoring mechanism thoroughly before starting the project.

Reference:

1. Baseline system: https://arxiv.org/abs/1803.05355/
2. FEVER challenge: http://fever.ai/

Project Option 2: Social Media Mining for Health Monitoring

Social media is a popular medium for the public to voice their opinions and thoughts on various health related topics. A recent Pew Research Center study says that nearly half of adults worldwide and two-thirds of all American adults use social networking on a regular basis. Due to the wealth of data available, researchers have been analyzing social media data for health monitoring and surveillance. However, social media mining for health issues is fraught with many linguistic variations and semantic complexities in terms of the various ways people express medication-related concepts and outcomes. This project requires processing imbalanced, noisy, real-world, and substantially creative language expressions from social media to extract and classify mentions of adverse drug reactions (ADRs) in tweets.

There are 4 tasks involved in this project:

Task 1: Automatic classification of adverse effects mentions in tweets

Classify the tweets reporting an adverse effect (AE) from those that do not. For each tweet, the data set contains: (i) the user ID, (ii) the tweet ID, and (iii) the binary annotation indicating the presence or absence of ADRs.

Dataset:

Training data: 25,672 (2,374 positive and 23,298 negative)
Evaluation data: approximately 5,000 tweets.
Evaluation metric: F-score for the ADR/positive class.
Link: https://data.mendeley.com/datasets/rxwfb3tysd/2/files/d2ae709b-c21e-420a-8de8-8b0f3abcfafe

Task 2: Extraction of Adverse Effect mentions

Identify the text span of the reported ADRs and distinguish ADRs from similar non-ADR expressions. ADRs are multi-token, descriptive, expressions, so this subtask requires advanced named entity recognition (NER) approaches. The data for this sub-task includes 2000+ tweets which are fully annotated for mentions of ADRs and Indications. This set contains a subset of the tweets from Task 1 tagged as hasADR plus an equal number of noADR tweets. Some tweets in the noADR subset were annotated for mentions of Indications to allow students to develop techniques to deal with this confusion class.
For each tweet, the data set contains: (i) the tweet ID, (ii) the start and (iii) end of the span, (iv) the annotation indicating an ADR or not and (v) the text covered by the span in the tweet.

Dataset:

Training data: 2,367 (1,212 positive and 1,155 negative)
Evaluation data: 1,000 (~500 positive, ~500 negative)
Evaluation metric: Strict and Relaxed F1-score, Precision and Recall
Link: https://data.mendeley.com/datasets/rxwfb3tysd/2/files/bc16c731-36b9-40fc-bfee-242495f7f139

Task 3: Normalization of adverse drug reaction mentions (ADR)
This is an end-to-end task, where the objective is to detect tweets mentioning an ADR and to map the extracted colloquial mentions of ADRs in the tweets to standard concept IDs in the MedDRA vocabulary (lower level terms). MedDRA (Medical Dictionary for Regulatory Activities) is the standard nomenclature for monitoring medical products, and includes diseases, disorders, signs, symptoms, adverse events or adverse drug reactions.

This task requires to understand the semantic interpretation of ADRs in order to map them to standard concept IDs. This task is likely to require a semi-supervised approach to successfully disambiguate ADRs. For each ADR mention, the publicly available data set contains: (i) the tweet ID, (ii) the start and (iii) end of the span, (iv) the annotation indicating an ADR or not, (v) the text covered by the span in the tweet and (iii) the corresponding ID of the preferred term in the MedDRA vocabulary.

Dataset:
Training data: 2,367 (1,212 positive and 1,155 negative)
Evaluation data: 1,000 (~500 positive, ~500 negative)
Evaluation metric: Strict and Relaxed F1-score, Precision and Recall
Link: https://data.mendeley.com/datasets/rxwfb3tysd/2/files/1959c8c0-1538-4bfe-96d6-d1bb84dc6a70

Task 4: Live Demonstration of an end-to-end system
Following up on the above tasks, students are required to build a website demonstrating an end-to-end system of ADR mentions in the tweets in the class. You are required to highlight the text span of ADR mentions/indications, annotation indicating ADR or not and overall sentiment of the tweet. In addition to using evaluation data as the data source for your website, you are also required to ingest more tweets on health-related topics.

Reference: https://www.aclweb.org/anthology/W19-3203.pdf

Project Option 3: Protest News

The goal of this project is to gain hands-on experience in developing practical solutions to societal problems based on web/text/social media mining. The specific event classes of interest are: global social unrest events, such as **demonstrations, marches**, and **protests** from 2019; and the countries to be focused on are:**India** and **South Africa**. This project seeks innovative solutions and approaches for discovering and combining multiple data sources.

Benchmark Dataset:

Your system should be evaluated against the benchmark dataset. The Armed Conflict Location & Event Data Project (ACLED) is a disaggregated conflict analysis and crisis mapping project. ACLED is the highest quality, most widely used, real-time data and analysis source on political violence and protest in the developing world. Practitioners, researchers and governments depend on ACLED for the latest reliable information on current conflict and disorder patterns. Refer to https://www.acleddata.com for more information and access to the data.

The project involves three major tasks:

Task 1: Data collection
As a first step, you are required to collect data on protests, marches and demonstrations that happened in 2019 in India and South Africa. For this task, you can make use of ACLED entries to find the related news articles for each event. Some of the libraries that you can use to scrape news articles are news-please, spacy, scrapy, newspaper. Please note that all the data should be in JSON format. Each group will be assigned a time span for data collection (say Jan-Mar, 2019 for group 1), hence all the groups are required to meet the instructors before starting the project (date and time to meet will be posted on Piazza).

Task 2: Event extraction and summarization
Your system should give detailed information about the events so that your users could understand the events by reading the summary. The following information should be provided for each event: event date, location, event type, parties involved, data sources, and a brief description. Daily based summarization should be generated on a timely manner, i.e. the delay of your system should be at most 24 hours.

Task 3: Live Demonstration of the end-to-end system
You are required to build a website to demonstrate your end-to-end system in the class. Your website must contain multiple dashboards showing insights into your data and visualizing events in both countries. The website should also show all the information related to the event: event date, location, event type, parties involved, data sources, and a brief description. Students are encouraged to make use of online tools and technologies to come up with detailed and insightful visualizations.

Evaluation Metrics:
For event detection - classification accuracy/precision/recall/F1
For filling in the slots (date, location, type etc) - Linear weights
For event summary/brief description - ROUGE score

What to submit

You should plan on preparing for the following:

1. **Project proposal**: Your proposal must contain the following sections:
   - Problem Statement - define the problem you are trying to solve, your objectives.

    - Literature Study - background reading on some state-of-the-art results, summarize them.
    - Dataset - details on the dataset, how the dataset is processed and adapted by your system.
    - Evaluation - which evaluation metrics are being used.
    - Proposed System - high-level architecture of your proposed system followed by a detailed explanation of each component of it.
    - Project Plan and Timeline - a clear plan of your project – who does what and the targets for each milestone.

2. **In-person presentation** of project plan, and plans for baseline system
3. **Midterm report** describing baseline system and initial evaluation results
4. **Final in-class presentation**
5. **Project report** in conference paper format

Grading

- **Milestone 1 (15%):** Project Proposal (week of Feb 25)

    - Literature Review
    - Project objectives
    - Data set, features to be implemented
    - Evaluation methodology
    - Project plan
    - Presentation of project plan

- **Milestone 2 (15%):** Baseline results (week of March 25)

- **Milestone 3 (30%):** Final Project Presentation (week of May 6th)
    - In class presentation
    - Project report (KDD paper format) to be submitted
    - All deliverables due by May 2

All project related discussion will be conducted through the piazza site for this course.