CSE 635: NLP and Text Mining

Spring 2021
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 rumour verification, chatbots, and social media mining related to Covid-19. 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 demo system. This may involve developing a 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: Stance Detection and Rumour Verification

Background: Rumours are rife on the web. False claims affect people's perceptions of events and their behaviour, sometimes in harmful ways. With the increasing reliance on the Web – social media, in particular – as a source of information and news updates by individuals, news professionals, and automated systems, the potential disruptive impact of rumours is further accentuated.

Within NLP research the tasks of stance classification of news articles and social media posts and the creation of systems to automatically identify false content are gaining momentum. The project requires to complete two of the RumourEval 2019 tasks, the first task is to classify/ detect stances of a tweet's reply thread and the second task is to verify the rumour introduced by the source tweet credibility.

Dataset: The data are structured as follows. Source posts introduce a rumour and may be true, false or unverified. These are accompanied by an ensuing discussion (tree-shaped) in which users support, deny, comment or query (SDCQ) the rumour in the source text.

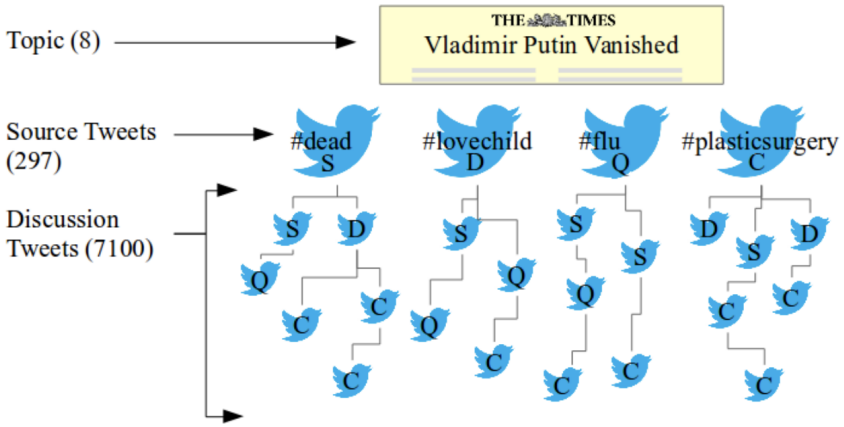

Figure 1: Structure of the first rumours corpus

As shown in Figure 1, for a particular topic there is a set of source tweets and for each source tweet, there is a thread of replies. Along with Twitter data, Reddit data is also available. The data files are arranged in this way:

Level1/2 files/dir	**Level 3 files/dir**	**Description**
<topic name>		
<source_tweet_id>		
	<context>	Additional data that can be used for classification.
	<replies>	Replies to the original tweet.

	<souce_tweet>	The source tweet.
	structure.json	The structure of the replies thread.
<train/dev>_key.json		This file contains all the labels.

The data distribution is shown in the table as follows:

	Supp.	Deny	Query	Com.	Total
Twitter Train	1004	415	464	3685	5568
Reddit Train	23	45	51	1015	1134
Total Train	1027	460	515	4700	6702
Twitter Test	141	92	62	771	1066
Reddit Test	16	54	31	705	806
Total Test	157	146	93	1476	1872
Total Task A	1184	606	608	6176	8574

Table 3: Task A corpus

	True	False	Unver.	Total
Twitter Train	145	74	106	325
Reddit Train	9	24	7	40
Total Train	154	98	113	365
Twitter Test	22	30	4	56
Reddit Test	9	10	6	25
Total Test	31	40	10	81
Total Task B	185	138	123	446

Table 4: Task B corpus

SDQC support classification. Example 1:

u1: We understand that there are two gunmen and up to a dozen hostages inside the cafe under siege at Sydney.. ISIS flags remain on display #7News **[support]**

- **u2:** @u1 not ISIS flags **[deny]**
- **u3:** @u1 sorry - how do you know its an ISIS flag? Can you actually confirm that? **[query]**
  - **u4:** @u3 no she cant cos its actually not **[deny]**
- **u5:** @u1 More on situation at Martin Place in Sydney, AU LINK **[comment]**
- **u6:** @u1 Have you actually confirmed its an ISIS flag or are you talking shit **[query]**

SDQC support classification. Example 2:

u1: These are not timid colours; soldiers back guarding Tomb of Unknown Soldier after today's shooting #StandforCanada PICTURE **[support]**

- **u2:** @u1 Apparently a hoax. Best to take Tweet down. [deny]
- **u3:** @u1 This photo was taken this morning, before the shooting. [deny]
- **u4:** @u1 I dont believe there are soldiers guarding this area right now. [deny]
  - **u5:** @u4 wondered as well. Ive reached out to someone who would know just to confirm that. Hopefully get response soon. [comment]
    - **u4:** @u5 ok, thanks. [comment]

Table 2: Examples of tree-structured threads discussing the veracity of a rumour, where the label associated with each tweet is the target of the SDQC support classification task.

Veracity prediction. Example 1:
u1: Hostage-taker in supermarket siege killed, reports say. #ParisAttacks LINK [true]
Veracity prediction. Example 2:
u1: OMG. #Prince rumoured to be performing in Toronto today. Exciting! [false]

Table 1: Examples of source tweets with veracity value

Dataset link: https://figshare.com/articles/dataset/RumourEval_2019_data/8845580

Task Definitions:

- **Subtask A - SDQC support classification:** given a source tweet, tweets in a conversation thread discussing the claim are classified as either supporting, denying, querying or commenting on the rumour mentioned by the source tweet. Success on this task supports success on task B by providing additional context and information; for example, where the discussion ends in a number of agreements, it could be inferred that human respondents have verified the rumour.
- **Subtask B - Veracity prediction**: the goal of subtask B is to predict the veracity of a given rumour. The rumour introduced by the source tweet that spawned the discussion is classified as true, false or unverified. Use of additional information from context and Subtask A results can significantly improve the performance. In addition to returning a classification of true, or false, a confidence score was also required, allowing for a finer-grained evaluation. A confidence score of 0 should be returned if the rumour is unverified.

Evaluation Metrics:

- **Subtask A:** Macro-averaged F1 is used to evaluate the classification performance.
- **Subtask B:** Again, macro-averaged F1 is used to evaluate the classification performance. For the confidence score, a root mean squared error (RMSE, a popular metric that differs only from the Brier score in being its square root) is to be calculated relative to reference confidence of 1.

References: RumourEval 2019 summary paper: https://www.aclweb.org/anthology/S19-2147.pdf

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. Recent studies indicate 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 multilingual tweets that report adverse effects (Task 2 in SMM4H 2020): This binary classification task involves distinguishing tweets that report an adverse effect (AE) of a medication (annotated as "1") from those that do not (annotated as "0"), taking into account subtle linguistic variations between AEs and indications (i.e., the reason for using the medication).

Dataset:

Training data: 25,672 tweets (2,374 "positive" tweets; 23,298 "negative" tweets)
Evaluation data: approximately 5,000 tweets.
Evaluation metric: F-score for the ADR/positive class.

tweet_id	user_id	class	tweet
3.4427E+17	809439366	0	depression hurts, cymbalta can help
3.4922E+17	323112996	0	@jessicama20045 right, but cipro can make things much worse...and why give bayer more of your money? they already screwed you once w/ essure
3.5142E+17	713100330	0	@fibby1123 are you on paxil .. i need help
3.2659E+17	543113070	0	@redicine the lamotrigine and sjs just made chaos more vengeful and sadistic.
3.4557E+17	138795534	0	have decided to skip my #humira shot today. my body's having hysterics, need time to simmer down #rheum
3.3259E+17	582163782	0	@needtobeskinny0 i was given 7 months worth of fluoxetine, at once. but my parents give it to me, i'm not trusted at all ;-;
3.4914E+17	1494435144	0	#bipolar meds think just #lithium for #monotherapy #lamotrigine for #rapidcycling #atypicalantipsychotics now revealing serious #sideeffects
3.403E+17	12926592	0	@shakeymike is feeling under the weather. i had to put out the trash. it's weird, i liked it. i even gave max his prozac myself. #imaboss
3.4197E+17	506346650	0	everyone's always upset in my house. damn, take a prozac
3.4898E+17	86229205	0	rt @neuronow: studies reinforce invokana(tm) (canagliflozin) (300mg) provides greater improvements in blood glucose than sit... http://t.co,Ä¶
3.4962E+17	536666835	0	i'd like to try venlafaxine.
3.4841E+17	252878361	0	rt @joshuagates: tip: my 5 item health kit for distant lands: cipro (gut), z pak (chest), pepto (stomach), purell (germs), advil (hangovers,Ä¶

Task 2: Automatic extraction and normalization of adverse effects in English tweets (Task 3 in SMM4H 2020): This task, organized for the first time in 2019, is an end-to-end task that involves extracting the span of text containing an adverse effect (AE) of a medication from tweets that report an AE, and then mapping the extracted AE to a standard concept ID in the MedDRA vocabulary (preferred terms). The training data includes tweets that report an AE (annotated as "1") and those that do not (annotated as "0"). For each tweet that reports an AE, the training data contains the span of text containing the AE, the character offsets of that span of text, and the MedDRA ID of the AE. For some of the tweets that do not report an AE, the training data contains the span of text containing an indication (i.e., the reason for using the medication) and the character offsets of that span of text, allowing participants to develop techniques for disambiguating AEs and indications.

Dataset:

Training data: 2,376 (1,212 positive and 1,155 negative)
Evaluation data: 1,000
Evaluation metric: Strict and Relaxed F1-score, Precision and Recall

tweet_id	begin	end	type	extraction	drug	tweet	meddra_cod	meddra_term
3.4389E+17	0	13	ADR	Restless arm	quetiapine	restless arms & legs! blood quetiapine :-/	10028006	motor restlessness
3.4389E+17	0	24	ADR	restless arm	quetiapine	restless arms & legs! blood quetiapine :-/	10038742	restless legs
3.4881E+17	0	42	ADR	Feels like on	seroquel	feels like one of those 5-cup #coffee days - but wait: i'm we	10015595	excessive daytime sleepiness
3.4891E+17	0	6	ADR	Bombed	olanzapine	bombed on olanzapine. work going to be tricky. 5 days of no	10070679	feeling stoned
3.4598E+17	0	6	ADR	Crying	effexor	crying randomly at nothing and everything. sigh. thank you #	10011469	crying
3.5293E+17	0	18	ADR	Allergic reac	lamotrigine	allergic reaction to #lamotrigine. feel free to share & d	10001718	allergic reaction
3.4487E+17	0	14	ADR	Almost vomi	lozenge	almost vomited on a zinc lozenge #yuk	10028822	nauseated
3.4299E+17	0	21	ADR	Sleeping my	quetiapine	sleeping my life away on #quetiapine. fine by me.	10041000	sleep excessive

Task 3: Classification of COVID19 tweets containing symptoms

(Task 6 in SMM4H 2021): Identifying personal mentions of COVID19 symptoms requires distinguishing personal mentions from other mentions such as symptoms reported by others and references to news articles or other sources. The classification medical symptoms from COVID-19 Twitter posts presents two key issues: First, there is plenty of discourse around news and scientific articles that describe medical symptoms. While this discourse is not related to any user in particular, it enhances the difficulty of identifying valuable user-reported information. Second, many users describe symptoms that other people experience, instead of their own, as they are usually caregivers or relatives of people presenting the symptoms. This makes the task of separating what the user is self-reporting particularly tricky, as the discourse is not only around personal experiences.

This task is considered a three-way classification task where the target classes are:

(1) self-reports, (2) non-personal reports, and (3) literature/news mentions.

Dataset:

Training data: 9,567 tweets
Evaluation data: 6,500 tweets
Evaluation metric: Micro F1 score

ID	Input	label
616	Pleurisy, dry cough, lung damage I can feel, slow/forgetful brain, tachycardia, extreme sleepiness, tired upon exertion, stinging toes, thirsty, bloat, and a partridge in a pear tree #COVID19	**self-report**
717	Okay. Either you had coronavirus and they still understand shit about it and what the testing actually means OR you've had some other insane illness that's caused a weird variety of symptoms (fatigue, cough) that they've failed to identify or treat for the last 3 months	**non-personal reports**
15	AI can now recognize COVID-19 from the sound of a cough. Based on a cellphone recording, machine learning models accurately detect coronavirus in a forced cough—even in people with no symptoms.Paper from Massachusetts Institute of Technology https://t.co/DTkS68uZVN	**Literature/news mentions**

Task 4: Classification of tweets self reporting potential COVID19 cases

(Task 5 in SMM4H 2021): This new binary classification task involves automatically

distinguishing tweets that self-report potential cases of COVID-19 (annotated as “1”) from those that do not (annotated as “0”). “Potential case” tweets include those indicating that the user or a member of the user’s household was denied testing for, symptomatic of, directly exposed to presumptive or confirmed cases of COVID-19, or has had experiences that pose a higher risk of exposure to COVID-19. “Other” tweets are related to COVID-19 and may discuss topics such as testing, symptoms, traveling, or social distancing, but do not indicate that the user or a member of the user’s household may be infected.

Dataset:

Training data: 7,181 tweets
Evaluation data: 10,000 tweets
Evaluation metric: F1 score

Tweet ID	Tweet Text	Class
12...497	I literally said I might have Coronavirus and isolated myself and today everyone is just touching me like I don’t wanna give u my corona so don’t touch me 😭😭😭	1
12...834	I'm at a loss as to how indifferent our government is towards banning flights, which ultimately plays a pivotal role in the spreading of this coronavirus.	0
12...369	All we know from my dr visit is I am sick. But I don’t have strep or the flu. But they also don’t have the means to test for Coronavirus	1
12...985	My office just put a work from home policy into effect until further notice. So that’s a thing. #coronavirus	0
12...672	I don't believe it would be responsible for me to go to SF. Do I think ~I~ will get coronavirus? No. Or if I do, I think it's survivable. Do I think I could spread it to others who might get sick if I do get it? Yes, and that would be bad.	0

Reference: https://healthlanguageprocessing.org/smm4h-sharedtask-2020/, https://healthlanguageprocessing.org/smm4h-2021/task-6/, https://www.aclweb.org/anthology/2020.smm4h-1.16/

Project Option 3: Conversational Question Answering (CoQA)

Humans gather information by engaging in conversations involving a series of interconnected questions and answers. For machines to assist in information gathering, it is therefore essential to enable them to answer conversational questions. We introduce CoQA, a novel dataset for building Conversational Question Answering systems. Our dataset contains 127k questions with answers, obtained from 8k conversations about text passages from seven diverse domains. The questions are conversational, and the answers are free-form text with their corresponding evidence highlighted in the passage. We analyze CoQA in depth and show that conversational questions have challenging phenomena not present in existing reading comprehension datasets, e.g., coreference and pragmatic reasoning. We evaluate strong conversational and reading comprehension models on CoQA. The best system obtains an F1 score of 65.4%, which is 23.4 points behind human performance (88.8%), indicating there is

ample room for improvement.

Benchmark Dataset:
CoQA is a large-scale dataset for building Conversational Question Answering systems. The goal of the CoQA challenge is to measure the ability of machines to understand a text passage and answer a series of interconnected questions that appear in a conversation. The datasets (both training and validation sets) are available here:
https://stanfordnlp.github.io/coqa/

Task : Question Answering using context
The ask is to train a model that can answer a series of questions from a user using the given comprehension. The questions posed by a user are conversational in nature .i.e. there might be coreferences within the questions. The model needs to find out the exact span of text from the given comprehension, which best answers the given question.

Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80. Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well. Jessica had . . .

Q_1: Who had a birthday?
A_1: Jessica
R_1: Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80.

Q_2: How old would she be?
A_2: 80
R_2: she was turning 80

Q_3: Did she plan to have any visitors?
A_3: Yes
R_3: Her granddaughter Annie was coming over

Q_4: How many?
A_4: Three
R_4: Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well.

Q_5: Who?
A_5: Annie, Melanie and Josh
R_5: Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well.

Figure 1: A conversation from the CoQA dataset. Each turn contains a question (Q_i), an answer (A_i) and a rationale (R_i) that supports the answer.

The Virginia governor's race, billed as the marquee battle of an otherwise anticlimactic 2013 election cycle, is shaping up to be a foregone conclusion. Democrat Terry McAuliffe, the longtime political fixer and moneyman, hasn't trailed in a poll since May. Barring a political miracle, Republican Ken Cuccinelli will be delivering a concession speech on Tuesday evening in Richmond. In recent ...

Q_1: What are the candidates **running** for?
A_1: Governor
R_1: The Virginia governor's race

Q_2: **Where**?
A_2: Virginia
R_2: The Virginia governor's race

Q_3: Who is the democratic candidate?
A_3: **Terry McAuliffe**
R_3: Democrat Terry McAuliffe

Q_4: Who is **his** opponent?
A_4: **Ken Cuccinelli**
R_4 Republican Ken Cuccinelli

Q_5: What party does **he** belong to?
A_5: Republican
R_5: Republican Ken Cuccinelli

Q_6: Which of **them** is winning?
A_6: Terry McAuliffe
R_6: Democrat Terry McAuliffe, the longtime political fixer and moneyman, hasn't trailed in a poll since May

Figure 2: A conversation showing coreference chains in color. The entity of focus changes in Q4, Q5, Q6.

Evaluation Metrics:
To evaluate your models, use the official evaluation script. To run the evaluation, use python evaluate-v1.0.py --data-file <path_to_dev-v1.0.json> --pred-file <path_to_predictions>. The evaluation model computes Macro F1 score for in domain and out of domain questions.
References: https://arxiv.org/pdf/1808.07042.pdf, https://stanfordnlp.github.io/coqa/

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.