Course Project Description - Type 1 (Undergraduate)

HLTH 453 / 619 – Fall 2025

This course project is mainly for undergraduate students enrolled in HLTH 453. In this project, you will create a Twitter-based vaccine hesitancy surveillance system. The underlying idea is that by monitoring the number of tweets that describe vaccine hesitancy issues or related consequences, it is possible to monitor trends at the provincial or country level. This kind of research is called "infodemics and infoveillance" research. This was particularly important when there was an urgent need to vaccinate a large portion of the Canadian population to protect against COVID-19.

There are two phases in this project: P1) tweet labelling, and P2) development of the surveillance system.

1 Tweet Labelling: Python Instructions

Please note: the following two steps of accessing Twitter data are out of date since Twitter became X. Please check how to access the X dataset. You can also use the public Twitter dataset for this assignment: https://github.com/echen102/COVID-19-TweetIDs

Each student is required to collect and manually label 100 tweets. There are two possible labels: *positive*, when the tweet is deemed related to vaccine hesitancy, and *negative* when the tweet is deemed to be not related. **Out of the 100 submitted tweets, exactly 10 must be positive**. A random sample (70%) of labelled tweets from the entire class will be made available for building your flu surveillance system, with 30% being held back for grading.

1. **Create a Twitter Developer (now X) account and obtain API keys**. Check out this post on the Twitter API Education Package. It may take a few hours or even a day for your account to be approved, so I recommend doing this step *before* you plan on sitting down and working with the code. Once your account has been approved, you will need to create a new App, and generate access and secret tokens for it. You'll also need your API key and secret key. *Do not share these keys, they are unique to your account.*

2. **Use the Tweepy package to collect live tweets.** The provided code in `twitter_streaming.py` will do most of the work here. To use the code, You first need to copy and paste your developer keys from Step 1 into the appropriate places in the code. You can then run it to connect to Twitter and start streaming live data. You can then work with the data as you see fit, but when you save it file **make sure to save full tweets, using the `status._json` attribute.**

   Note that the code creates a `filter` that constrains the search to certain keywords. This filter is already configured with selected search terms and to only capture tweets in English; **you should not modify this filter**.

   Warning: This code runs indefinitely so you will need to end it manually by pressing the stop button in Juypter or Ctrl+c in the command-line. You may also have to collect tweets for a reasonably long period of time to see true positives, since the vast majority of tweets *do not* relate to vaccine hesitancy.

3. **Manually label the tweets you collected.** Use the `TweetTagger` class you created for Assignment 1 to facilitate this labelling task. Use the `save()` method to produce the two files to be submitted: `tweets.txt` and `labels.txt`. As described in Assignment 1, the tweets and labels in the two files should be in the same order. Also recall that you may use `trim()` to remove excessive positives and negatives.

Resources:

- Twitter Developer Website
- Tweepy Documentation

2 Tweet Labelling: Coding Instructions

It's important that when we create a shared dataset of labelled tweets that we are all labelling the same thing. Our classifiers will depend on a reliable "ground truth" dataset, from which they can extrapolate and accurately predict unseen tweets. To ensure that we're all on the same page when creating our separate datasets, we will use a codebook developed by Blankenship et al. [2018]. In particular, they provide some definitions and examples for *pro-vaccine*, *neutral*, and *anti-vaccine* tweets.

1. **pro-vaccine**, e.g. "Equine #Influenza #Vaccine Remains Effective Against Mutated Virus Study Shows - TheHorse.com : http://bit.ly/c4eRWu"
2. **neutral**, e.g. "#H1N1 #vaccine still available recommended - Zanesville Times Recorder: http://bit.ly/aJQyIY"
3. **anti-vaccine**, e.g. "RT @bengoldacre "Give children #vaccine that we said would kill them": Daily Mail joy by @PrimlyStable http://dlvr.it/DKJRm #health"

For this assignment, you should label 'pro-vaccine' and 'neutral' tweets as *'negative'*, and 'anti-vaccine' tweets as *'positive'*.

Blankenship et al. [2018] also identified sub-categories of anti-vaccine tweets. For the purposes of creating this dataset, you do not need to consider those subcategories as separate labels. However, they do provide some useful guidance in determining if a tweet should be considered to be 'anti-vaccine':

1. Perceived harmful risks, side effects, and/or deaths caused by vaccines (eg, autism, seizures, fatalities)
   (a) Example: "Dr. #Russell #Blaylock on #Alex #Jones Tv 3/3:Harmful Side-Effects of The #Swine Flu #Vaccine! http://goo.gl/fb/e4ccf"
   (b) Example: "RT @CassieSunset: The true story of SV40 the cancer-causing virus hidden in polio vaccines http://bit.ly/lCOkIi #cancer #vaccine"
2. Distrust of government, pharmaceutical companies, and/or scientists (includes supporting organizations such as the Bill & Melinda Gates Foundation)
   (a) Example: "so anyhow that is just one example of a mindless #pharma #shill trying to preach the #vaccine #gospel ... #mindless #sheeple ... go line up"

3. Miscellaneous

   (a) Example: "What is this all about? School Expels Child Over #Vaccine That Even the Doctor Refuses to Give http://t.co/5cFuHDOsqw #health #preppertalk"

   (b) Example: "The #vaccine empire has collapsed. http://t.co/z3Yp7tXI"

   (c) Example: "#Whooping Cough Outbreak at California School Among Vaccinated Shows #Vaccine:http://t.co/LsGm3WBwA6 #WhoopingCough #health"

You may find it useful to review their work when creating your surveillance system, and in determining which features are most useful to consider when examining tweets. You can find their supplementary materials at http://www.thepermanentejournal.org/files/2018/17-138-Suppl.pdf

3 Create Your Surveillance System

The actual surveillance system should be coded in a Python file called `twitter_surveillance.py`. Its main functions are to:

1. Load your trained classifier from a file named `twitter_classifier.pkl`;
2. Read tweets from a file named `tweets.txt`;
3. Extract features from each tweet, and using the extracted features, classify each tweet as either *positive* or *negative*;
4. Write the corresponding labels to a file called `labels.txt`; and
5. Save a plot of daily percentages of positive tweets to a file called `plot.png`

Your tweet classifier must be a machine learning model trained using `scikit-learn` or `NLTK` — you can choose freely between them. 70% of all labelled tweets from the class will be made available for training, whereas the remaining 30% will be held back to evaluate your classifier. The generated labels should be in the same order as the tweets in `tweets.txt`, and saved to file in the same way as the `save()` function in Assignments 1, Question 3.

Your surveillance system should save its generated plot to a file called `plot.png`. To test your surveillance system prior to submission, you should have tweets from different days, which the training data from the class should provide. You can use the time stamps associated with each tweet to determine when it was created. If you want to get started early, you can work with the tweets you collected and labelled yourself.

4 Grading

This project is worth 50% of your total grade in the course. There are three components in this evaluation.

1. P1: [15%] Your 100 labelled tweets should be submitted to Marmoset by **11:55pm on Sunday, September** 28^{th}. They will be graded using Marmoset, and a manual check of a random sample of tweets.

2. P2: [15%] Surveillance system in Python. Your surveillance system will be evaluated in Marmoset, a grade of zero will be given for this component if running the submitted code results in errors. Classification performance will be evaluated using the test data that was held out from the labelled tweets collected from the class (30% of all labeled tweets). This part of the project is **due at 11:55pm on Sunday, December 7^{th}**.

3. P3: [20%] Final Exam: Presentation and oral exam (i.e. explaining of your project design, logic and code) - this is in-person 1-1 exam, and the sign-up sheet will be posted in reading week.

References

Elizabeth B Blankenship, Mary Elizabeth Goff, Jinging Yin, Zion Tsz Ho Tse, King-Wa Fu, Hai Liang, Nitin Saroha, and Isaac Chun-Hai Fung. Sentiment, contents, and retweets: a study of two vaccine-related twitter datasets. *The Permanente Journal*, 22, 2018. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6004971/.