**Title**:Bot or Not: An Exploration of Fake Yelp Review Detection

**Group Members:** Lauren D’Arinzo (lhd258), Elizabeth (Lisa) Combs (eac721), Angela Marie (Amber) Teng (at2507) - *responsible for gradescope submission*, Paula Kiatkamolwong (kk4158), Steven Dornberg (dornbs01)

**Summary of Plans**:

*Chosen Project*

The team has selected to complete project 1: fake review detection.

* *Text preprocessing* (i.e. HW1 preprocessing function + slides) including: word stemming, lemmatization, remove stopwords, group common words.
* *Generate additional features* including but not limited: n-gram, length of review, number of reviews per user\_id, etc.

*Proposed Approach*

1. Baseline models: Naive Bayes and Decision Tree
2. Logistic Regression, SVM
3. Random Forest, XGBoost, AdaBoost

*Suggested Experiments*

1. Hyperparameter tuning for the optimal algorithm
2. Oversampling, undersampling
3. With additional time, we could test sentiment analysis on the reviews.

**Title**:Predicting the Severity of COVID-19 Cases in the Philippines

**Group Members:** Lauren D’Arinzo (lhd258), Elizabeth Combs (eac721), Angela Marie (Amber) Teng (at2507) - *responsible*, Paula Kiatkamolwong (kk4158), Steven Dornberg (dornbs01)

**Summary of Plans**:

*Chosen Project:* The team selected to complete project 3: “fighting COVID-19”. We will use a [dataset](https://docs.google.com/spreadsheets/d/16g_PUxKYMC0XjeEKF6FPUBq2-pFgmTkHoj5lbVrGLhE/edit?fbclid=IwAR29_EijSX1_bTS1oNW6nuYNjMvoN-HcP44PCEoM13JSfXD0gf1Uu2D2XoE#gid=0) from COVID-19 Philippines, which is a live source gathered by Filipino data scientists, and obtained from the Philippine Department of Health (DOH) records of COVID-19 reported cases. The dataset was compiled and is currently managed by the Philippine Data Science [Group](https://www.facebook.com/datasciencephilippines). We chose this dataset because it contains [patient-level data](https://public.tableau.com/profile/rozauro.cordon?fbclid=IwAR1EaH7YlwhyylXkE4reI3wwBlbayTcyrl1W3KbxDLTnRnHo5BU-Mc_6P0I#!/vizhome/COVID-19PHTracker/COVID-19?publish=yes), which would allow us to answer the research question: Given a presumed COVID-19 diagnosis and features like age, travel history, epidemiological virus links etc, can we predict a person’s health status in relation to COVID-19? Particularly, will they have a mild or severe case? The dataset also contains opportunities for generating additional features including the length of time between on-set of symptoms and their admittance to a hospital.

*Proposed Approach:* We propose to use multiclass classification to solve this question. Our baseline models will consist of Naive Bayes and Decision Tree, which are naturally extensible to the case of having more than two classes. We also plan to test whether using SVM, K-nearest neighbors, and neural networks improves our error rate.  
*Suggested Experiments:* In finetuning our model, we will perform hyperparameter tuning for the optimal algorithm. We can also use oversampling, undersampling techniques to better address our imbalanced dataset. Finally, with additional time, we could add additional data from other countries to our analysis, including data from this [source](https://docs.google.com/spreadsheets/d/e/2PACX-1vQU0SIALScXx8VXDX7yKNKWWPKE1YjFlWc6VTEVSN45CklWWf-uWmprQIyLtoPDA18tX9cFDr-aQ9S6/pubhtml#). Another dataset we could review is news articles from the original source data that could allow us to add more features.

*References:*

1. <https://public.tableau.com/profile/rozauro.cordon?fbclid=IwAR1EaH7YlwhyylXkE4reI3wwBlbayTcyrl1W3KbxDLTnRnHo5BU-Mc_6P0I#!/vizhome/COVID-19PHTracker/COVID-19?publish=yes>
2. <https://news.abs-cbn.com/news/03/15/20/list-health-departments-list-of-confirmed-covid-19-cases?fbclid=IwAR29_EijSX1_bTS1oNW6nuYNjMvoN-HcP44PCEoM13JSfXD0gf1Uu2D2XoE>
3. <https://www.facebook.com/groups/datasciencephilippines/?post_id=897213460732154>
4. <https://www.doh.gov.ph/covid-19/case-tracker>
5. <https://docs.google.com/spreadsheets/u/1/d/e/2PACX-1vQU0SIALScXx8VXDX7yKNKWWPKE1YjFlWc6VTEVSN45CklWWf-uWmprQIyLtoPDA18tX9cFDr-aQ9S6/pubhtml#>
6. <https://en.wikipedia.org/wiki/Multiclass_classification#Naive_Bayes>

*Other Data Sources:*

1. <https://covidtracking.com/data/>
2. <https://www.tableau.com/covid-19-coronavirus-data-resources>
3. <https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases>
4. <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/>
5. <https://www.doh.gov.ph/covid-19/case-tracker>
6. <https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge>
7. <https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data>
8. <https://www.safegraph.com/dashboard/covid19-commerce-patterns?fbclid=IwAR2vMqEtNSincgmCaLSde-eZ2lLxTCoKbftXHXjhIhe8XepIXudXZQaPyVg>
9. <https://github.com/beoutbreakprepared/nCoV2019/blob/master/source_list.csv>
10. <https://towardsdatascience.com/5-datasets-about-covid-19-you-can-use-right-now-46307b1406a>
11. <https://docs.google.com/spreadsheets/u/1/d/e/2PACX-1vQU0SIALScXx8VXDX7yKNKWWPKE1YjFlWc6VTEVSN45CklWWf-uWmprQIyLtoPDA18tX9cFDr-aQ9S6/pubhtml#>

Milestones:

1. CodaLab & Collab File set up
2. Data cleaning + text preprocessing
   1. Diff preprocessing (bigram, capitalization, stemming etc)
3. EDA
4. Baseline
   1. Different sampling methods
      1. 90/10
      2. 50/50
5. Hyperparameter Tuning
6. Experiment
   1. Ensemble?
7. Report Writing

Questions to ask the Group:

* What is the difference between status and health status?
* Would it be okay to add a column with the link to the news article source per patient?
* How often is the dataset updated?
* What is the difference between longitude latitude and residence longitude residence latitude?

To-Do:

1. Codalab
2. Collab File

# 

# Meeting Notes 03/24/2020:

[**https://docs.google.com/document/d/10ZalnCfSVQeUGrh8ZCtyaNgyWYg-H1W6/edit**](https://docs.google.com/document/d/10ZalnCfSVQeUGrh8ZCtyaNgyWYg-H1W6/edit)

[**https://worksheets.codalab.org/worksheets/0x33171fbfe67049fd9b0d61962c1d05ff**](https://worksheets.codalab.org/worksheets/0x33171fbfe67049fd9b0d61962c1d05ff)

[**https://www.kaggle.com/nicodds/rome-wasn-t-built-in-a-day-spotting-fake-reviews**](https://meet.google.com/linkredirect?authuser=2&dest=https%3A%2F%2Fwww.kaggle.com%2Fnicodds%2Frome-wasn-t-built-in-a-day-spotting-fake-reviews)

## **Project Proposal**

Groups must upload a one page pdf file on Gradescope containing:

* Title
* Group
  + Name and NetID of each member.
  + Member responsible for uploading submissions.
* Summary of Plans
  + Which project you have chosen
  + Proposed Approach
  + Suggested Experiments

## **Project report**

Groups must upload a pdf on Gradescope describing their approach and results.

1. Title
2. Group Members
   1. Name and NetID of each member.
   2. Member responsible for uploading submissions
3. Introduction
   1. Description of Problem
   2. Approach
   3. Summary of results/contribution
4. Approach
   1. Describes the details of your approach
5. Experiments
   1. Description of Datasets
   2. Baselines or other approaches for comparison
   3. Explanation of Results
   4. Error analysis
6. Discussion
   1. Evaluation of Findings
   2. Possible Next Steps

NOTE: Only the *member responsible for uploads* needs to upload the pdf file. In other words, each group should have only one pdf file uploaded on Gradescope.

## **Evaluation**

The final project will be graded based on three main aspects:

1. adherence to guidelines
2. quality of the report and final results
3. efforts to model development, e.g. a novel algorithm will be recognized

A final report should:

1. clearly state the problem, pointing to hurdles and issues to solve it;
2. clearly present the methodology employed to solve the problem, pointing out:
   * 1. the data sets used
     2. the methods employed to (if necessary) handle missing data, transform data, combine data, etc.
     3. the algorithms involved in the solution, as for example, SVM for classification, DBScan for clustering, etc.
     4. present and discuss the results, highlighting the strengths and weaknesses of the proposed methodology
     5. make some conclusion, emphasizing whether the chosen approach was success and, if not, why.

## **Template**

Below are guidelines on how to write-up your report for the final project. Not all of the

comments may not be relevant to every project. However, please use it as a general guide in

structuring your final report. A “standard” experimental machine learning paper consists of the

following sections:

**1. Introduction**

Motivate and abstractly describe the problem you are solving and how you are addressing it. What is the problem? Why is it important? What is your basic approach? A short discussion of how it fits into related work in the area is also desirable (optional for this assignment). Summarize the basic results and conclusions that you will present.

**2. Related Work**

This section is optional. If in working on your project you came across other papers tackling the same or a similar problem, cite and describe the related work: What is their problem and method? How is your problem and method different? Why might your approach be better? How does your work fit in the bigger picture?

**3. Problem Definition and Algorithm**

**3.1 Task**

Precisely define the problem you are addressing (i.e. formally specify the inputs and outputs).

**3.2 Algorithm**

Describe in reasonable detail the algorithm(s) you are using to address this problem. A pseudocode description of the algorithm(s) you are using is frequently useful. Trace through a concrete example, showing how your algorithm processes this example. The example should be complex enough to illustrate all of the important aspects of the problem but simple enough to be easily understood. If possible, an intuitively meaningful example is better than one with meaningless symbols. Your description of the algorithm should include what assumptions if any you are making about the data, and also what parameters or design choices need to be made (the consequences of these choices should then be explored in detail in the experimental evaluation).

**4 Experimental Evaluation**

**4.1 Data**

Describe the data sets that you use in your experimental evaluation. If you do any feature pre-processing, this is the place to describe it.

**4.2 Methodology**

Describe the experimental methodology that you used. What are the criteria that you are using to evaluate your method? What specific hypotheses does your experiment test? How did you do training/validate/test splits? Comparisons to competing methods that address the same problem are particularly useful.

**4.3 Results**

Present the quantitative results of your experiments. Graphical data presentation such as graphs and histograms are frequently better than tables. What are the basic differences revealed in the data? Are they statistically significant?

**4.4 Discussion**

Is your hypothesis supported? What conclusions do the results support about the strengths and weaknesses of your method compared to other methods? How can the results be explained in terms of the underlying properties of the algorithm and/or the data.

**5 Conclusions**

Briefly summarize the important results and conclusions presented in the paper. What are the most important points illustrated by your work? If you were to continue working on the project, what are the interesting areas for future work? What are the major shortcomings of your current method? For each shortcoming, propose additions or enhancements that would help overcome it.

**6 Bilbiography**

Be sure to include a standard, well-formated, comprehensive bibliography with citations from the text referring to previously published papers in the scientific literature, resources, or code that you utilized or referenced during your project.