# CSC 219-01 Machine Learning (Fall 2023)

# Project 3: Multi-Modal Deep Learning for Fake News Detection

**Due at 3:00 pm, Monday, October 30, 2023**

**Demo: class time, Monday, October 30, 2023**

## 1. Problem Formulation

In this project, we will build a multi-modal deep learning framework for fake news (tweets) detection on social media by using both textual (tweets) and visual inputs (images). This proposed model aims to deal with the automatic detection of manipulation and misuse in Web multimedia content, which lays the basis for a future generation of tools that could assist media professionals in the process of social media information verification.

## 2. Dataset

We will use a large-scale multimodal fake news dataset consisting of over 1 million samples containing text, image, metadata, and comments data from a highly diverse set of resources. Each data sample consists of multiple labels, allowing users to utilize the dataset for 2-way, 3-way, and 6- way classification. See the basic stats below:

| Dataset Statistics | |
|---|---|
| Total samples | 1,063,106 |
| Fake samples | 628,501 |
| True samples | 527,049 |
| Multimodal samples | 682,996 |
| Subreddits | 22 |
| Unique users | 358,504 |
| Unique domains | 24,203 |
| Timespan | 3/19/2008 - 10/24/2019 |
| Mean words per submission | 8.27 |
| Mean comments per submission | 17.94 |
| Vocabulary size | 175,566 |
| Training set size | 878,218 |
| Validation set size | 92,444 |
| Released test set size | 92,444 |
| Unreleased set size | 92,444 |

The goal of this project is to build a 2-way classification deep learning model with two inputs and one output using TensorFlow Functional API.

Download the dataset here: https://github.com/entitize/Fakeddit

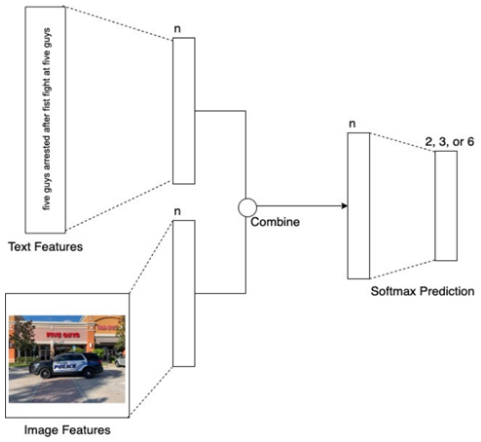
In this project, we will only focus on multimodal samples (samples that have both text and image). So only use the samples in the ***multimodal_only_samples*** folder. Use the ***clean_title*** column to get filtered **text data.**

Use ***multimodal_train.tsv*** for training and ***multimodal_validate.tsv*** for test. To speed up training, you may use Google Colab. Feel free to work only on a **random subset** of each document if it takes too much time for training the model.

| author | clean_title | created_utc | domain | hasimage | id | image_url | linked_submission_id | num_comments | score | subreddit | title | upvote_ratio | 2_way_label | 3_way_label |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| trustbytrust | stargazer | 1425138660 | | TRUE | cozywbv | http://i.imgur.com/B | 2xct9d | | 3 | psbattle_artwork | stargazer | | 0 | |
| | yeah | 1438172995 | | TRUE | ctk61yw | http://i.imgur.com/JF | 3f0h7o | | 2 | psbattle_artwork | yeah | | 0 | |
| chaseoes | pd phoenix car thief | 1560491936 | abc15.com | TRUE | c0gl7r | https://external-preview.redd.it/1A2_4Vwg | | 2 | 16 | nottheonion | PD: Phoenix car thie | 0.89 | 1 | |
| SFepicure | as trump accuses ira | 1560605928 | nytimes.com | TRUE | c0xdqy | https://external-preview.redd.it/9BKRcgvac | | 4 | 45 | neutralnews | As Trump Accuses Ir | 0.78 | 1 | |
| fragments_from_Wor | believers hezbollah | 1515138811 | i.imgur.com | TRUE | 7o9mx | https://external-preview.redd.it/rbwXHncnjY | | 40 | 285 | propagandaposters | "Believers" - Hezbolli | 0.95 | 0 | |
| SovietTurnipFarmer | the rise of italian fas | 1555334618 | i.redd.it | TRUE | bdfxf1 | https://preview.redd.it/jpapt6aufs21.jpg?w | | 2 | 10 | fakehistorypom | The rise of Italian fa | 0.92 | 0 | |
| HR_Paperstacks_402 | trumps pick to lead i | 1525130910 | washingtonpost.com | TRUE | 8g3xtm | https://external-preview.redd.it/4-bDBCtwD | | 1 | 17 | neutralnews | Trump's pick to lead | 0.75 | 1 | |
| lightblueyoshi | my friend midflip | 1401165057 | i.imgur.com | TRUE | 26kwdf | https://external-preview.redd.it/Kz4CnSWP | | 8 | 7 | photoshopbattles | My friend mid-flip | 0.82 | 1 | |
| pakattack461 | look at this cool ipho | 1420945357 | i.imgur.com | TRUE | 2s0xuj | https://external-preview.redd.it/m09Q0aKB | | 4 | 55 | misleadingthumbnail | Look at this cool IPh | 0.79 | 0 | |

Note that the entire image file is about 106GB. I strongly recommend you using the URLs (the ***image_url*** column in the tsv dataset files) to download ONLY the images you need. You may use the script provided. Copy ***image_downloader.py*** to the same folder where you downloaded the tsv files.

## 3. Model Design




Hint: To load images from disk, you may use *tf.keras.utils.image_dataset_from_directory()*. Please check these two great tutorials:
https://androidkt.com/use-image-dataset-from-directory-with-and-without-label-list-in-keras/
https://notebook.community/tensorflow/docs-l10n/site/en-snapshot/tutorials/load_data/images

## 4. Requirements

- To encode the text, use pre-trained GloVe embedding. Check our lab tutorial on transfer learning using GolVe.
- Make sure you match images with their corresponding texts.
- Use EarlyStopping when training neural networks using Tensorflow.
- NO need to do any hyper-parameter tuning.
- Print out Recall, Precision, and F1 score on test data for real and fake news, respectively.

## 5. Grading Breakdown

You may feel this project is described with some certain degree of vagueness, which is left on purpose. In other words, **creativity is strongly encouraged**. Your grade for this project will be based on the soundness of your design, the novelty of your work, and the effort you put into the project.

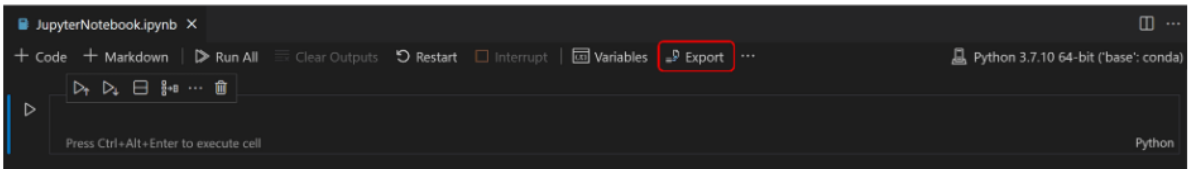
Use the evaluation form on Canvas as a checklist to make sure your work meet all the requirements.

## 6. Teaming

Students must work in teams of at most 3 people. Think clearly about who will do what on the project. Normally people in the same group will receive the same grade. However, the instructor reserves the right to assign different grades to team members depending on their contributions. So you should choose partner carefully!

## 7. Deliverables

**(1) The HTML version of your notebook that includes all your source code.** In VS Code, you can export a Jupyter Notebook as an HTML file. To export, select the Export action on the main toolbar. You'll then be presented with a dropdown of file format options.



**(2) Your report in PDF format,** with your name, your id, course title, assignment id, and due date on the first page. As for length, I would expect a report with more than one page. Your report should include the following sections (but not limited to):

- Problem Statement
- Methodology
- Experimental Results and Analysis
- Task Division and Project Reflection
- Additional Features

**In the section "Task Division and Project Reflection", describe the following:**

- who is responsible for which part,
- challenges your group encountered and how you solved them
- and what you have learned from the project as a team.

**In the section "Additional Features", you describe and claim credit for additional features.**

**To submit your notebook and report, go to Canvas "Assignments" and use "Project 3".**

All the deliverables must be submitted **by team leader** on Canvas before

**3:00 pm, Monday, October 30, 2023**

NO late submissions will be accepted.

## 8. Possible Additional Features (10 points each feature)

- Can you try transfer learning for the CNN input channel? https://keras.io/api/applications/ There are some models you may want to consider (try at least one model as listed here)
  - VGG16
  - ResNet50
  - MobileNetV2

- Use contextual word embedding like BERT to represent the text for the LSTM input channel. Here is a nice introduction: https://mccormickml.com/2019/05/14/BERT-word-embeddings-tutorial/ You may use any high-level implementation like the following:
  - https://www.sbert.net/docs/quickstart.html
  - https://github.com/llSourcell/bert-as-service

## 9. In-class Presentation.

On the due day, each team has 5 minutes to present your work in the class. Explain your solutions by referring to your notebook. You do not have to prepare the PowerPoint slides for your presentation.