MAST30034 Applied Data Science – Final Project Proposal

Title: Fake News Detection.

Team: Charlotte Williams (834810), Krystal Dowling (914430), Yifan Luo (926261), Laura Heard (994900).

Tutorial: Hossein Alipour, Thursday 11am.

Datasets:

Given the limitations of available datasets for fake news detection we have decided to combine multiple datasets of classified real and fake news. This will ideally give us a better analysis for the data by limiting the effect and bias of a singular data source and increasing the variability and size of the data to thus hopefully improve the accuracy of the classifiers.

Some reputable datasets that we are planning to use for our fake news detection classifier:

Fake News (Kaggle, <https://www.kaggle.com/c/fake-news/discussion>)

Getting Real about Fake News (Kaggle, <https://www.kaggle.com/mrisdal/fake-news>)

The Signal Media One Million News Articles Dataset (Signal, <https://research.signal-ai.com/newsir16/signal-dataset.html>)

PolitiFact/ Buzzfeed Real and Fake News Context (Kaggle, <https://www.kaggle.com/mdepak/fakenewsnet?select=PolitiFact_real_news_content.csv>)

# Source based Fake News Classification

# (Kaggle, https://www.kaggle.com/ruchi798/source-based-news-classification)

Task/s: Our task will be to classify news articles as either fake or real news, using sentiment analysis/ natural language processing. Given the success of our initial classification system and if time allows, we may consider exploring further extensions of the project including implementing a GAN to trick our own network.

SENTENCE ABOUT WHY WE CHOSE THIS TOPIC

**Method**

Pre-processing:

* Removing stop words, punctuation
* Merge datasets, concatenate.
* Ensuring the classifier can’t easily determine whether it is from a specific news source, based on text formatting, or name included in the text.
* Remove articles which are: Non-English, has missing values, outliers (contextual), noisy data.
* Split training/validation/testing
  + Cross-validation

Feature Engineering/ selection:

* GloVe embeddings
* Word Vectorizer
* Doc2Vec
* Bag-of words
* Chi-squared, mutual information test to determine significance of attribute

Models:

Want to contrast neural network models against non-neural models

* Baseline model: Logistic Regression
* Random forest
* Neural network
  + LSTM (Long Short-Term Memory)
  + Transformer
  + Convolutional Neural Network <- longer text classification

Analysis:

* Impact of news source/ location/ datetime.
  + Map plot visualisation.
  + Heatmap correlation between attributes.
  + Line graph for timeline.
* Title or body better for prediction.
* Best words for fake news.
  + Word clouds for popular words.
* Sort news into topics ie. Political, health etc.
  + Bar graph

Evaluation:

* Accuracy, precision, recall
* Bias of model
* Error rate
* Confusion matrix

Extensions:

* Stance detection: disconnect between body and title
* GAN; can we create fake news that can trick a human?