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 news articles classified as real or 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)
- FakeNewsNet (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

We chose our topic of fake news detection as we believe it an interesting, challenging, and relevant topic in the age of political warfare and COVID-19 as readers struggle to distinguish real news from fake (Asr & Taboada, 2019; Nyilasy, 2020). This problem is evidenced within data and thus can be explored through and would ideally benefit from analysis and classification.

Therefore, our task will be to classify news articles as either fake or real news, using sentiment analysis and natural language processing.

METHOD

For the method, we have outlined the processes we intend to take as a group in achieving our task and have provided ideas of steps we may implement to fulfil these. While we have not yet decided on an explicit and concise methodology at this point, we have demonstrated some of the methods we are considering employing, dependant on our future research on such methods, their appropriateness for the task, and our ability to implement them.

Pre-Processing and Data Cleaning

First, we plan to pre-process and clean the data to best be utilised by our analysis.

- Format datasets to concatenate into one
- Removing stop words, punctuation for NLP
- Remove articles/instances which are/contain: non-English text, missing values, outliers (contextual), noisy data.
- Split the data into training/validation/testing sets.
    - Implemented via cross-validation

- 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.

Feature Engineering and Selection

We will use one of the following methods of feature generation in order to vectorise the words in the text so they can be processed by our model. GloVe embeddings appear to be the most effective since it takes into account both global and local sequencing. Feature selection will then be used to refine our model with the most important attributes to use for classification.

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

Model Fitting and Classification

Next, we plan to create and compare classifiers in fake news detection. We plan to develop a baseline model for analysis and then contrast a deep learning model against a machine learning approach. For our machine learning classifier, we are considering logistic regression, naïve-bayes, and random forests as classifiers that have been utilised in other research as being appropriate for the problem (Chauhan, 2019).

For the deep learning classifier, we will be using a neural network. From the research, we are considering Recurrent Neural Networks (RNN), Transformer, and Convolutional Neural Network. Since the data we are utilising exists as a raw text sequential format, we feel RNN is an appropriate model to use as it is cited as being especially good at classifying sequential data (Elvis, 2018). RNN recursively applies a computation to every instance of an input sequence conditioned on the previous computed results. An RNN has the capacity to memorise the results of previous computations and use this information to inform the current computation. RNN's input are typically some sort of embedding for textual data. RNN's are used widely in many NLP applications such as in semantic analysis and semantic matching - match a message to candidate response in dialogue systems (Elvis, 2018). Similarly, convolutional neural networks have been cited as being more appropriate for longer text classification (Asr & Taboada, 2019).

Evaluation

Furthermore, we intend to use classical evaluation methods, outlined below, to evaluate our classifiers for analysis and comparison.

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

Analysis

For our analysis, we have outlined some facets and interactions within the data that we plan for analysis and discussion, with possible visualisations described.

- Comparison of models to evaluate, analyse, and discuss which was best and possibly why
- 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 i.e. political, health etc.
    - Bar graph

Possible Extensions

Our research on the topic has found various interesting higher-level methods of analyses of fake news data. Given the success of our initial classification system and if time allows, we may consider exploring further extensions of the project.

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

BIBLIOGRAPHY

Asr, F. T. & Taboada, M., 2019. Big Data and quality data for fake news and misinformation detection. *Big Data & Society,* 6(1).

Chauhan, K., 2019. *Exploratory Analysis and fake news classification on Buzzfeed News,* s.l.: Kaggle.

Elvis, 2018. *Deep Learning for NLP: An Overview of Recent Trends.* [Online]
Available at: https://medium.com/dair-ai/deep-learning-for-nlp-an-overview-of-recent-trends-d0d8f40a776d
[Accessed 15 September 2020].

Nyilasy, G., 2020. Fake News in the Age of COVID-19. *Inside Business*, 10 April.