Udacity – Capstone Proposal

Domain Background

The Capstone Proposal I am submitting to your attention is related to Fake News detection.

In an era where the impact of mass media and news is critical on the way of acting and reasoning of the population, the large amounts of data make it really difficult to be able to discern between fake and true news. I would like to personally address this task as I'm really interested in Natural Language Processing.

Problem Statement

Given an article, I want to try to build an algorithm which is able to determine whether a news article is true or fake (within possibility). The problem can be seen as a binary classification problem where the task is to predict a class label (0 for True news, 1 for Fake news) given the text, title and subject of the article.

Datasets and inputs

To try and solve this task, I will use a dataset taken from Kaggle, namely "Fake and real news dataset": https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset

The dataset features a list of articles, together with the subject of the article and its title categorized as Fake or True. The data is almost evenly distributed, with 20826 True articles and 17903 Fake articles divided in two files.

The dataset cites the following articles for acknowledgments:

- Ahmed H, Traore I, Saad S. "Detecting opinion spams and fake news using text classification", Journal of Security and Privacy, Volume 1, Issue 1, Wiley, January/February 2018.
- Ahmed H, Traore I, Saad S. (2017) "Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. In: Traore I., Woungang I., Awad A. (eds) Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments. ISDDC 2017. Lecture Notes in Computer Science, vol 10618. Springer, Cham (pp. 127-138).

Solution Statement

To try solving this problem, I might try several models among Machine Learning (Tf-Idf with a Binary Classifier) and Deep Learning (I could try Embeddings with LSTMs with a final Dense layer or Transformers) to try and see which performs better and is not too costly to use in terms both of training and time and inference time.

Evaluation Metrics

The evaluation metrics could involve accuracy, as the classes are pretty well balanced, but I could also try giving more importance to Precision, for example by labelling as 0 the True news and as 1 the Fake to see the detection power of the algorithm.

Benchmark Model

In the second article (2017) the authors report a Linear Support Vector Machine with an accuracy of 92%, so trying to achieve at least the same result is desirable.

Project Design

The project will be developed on Amazon SageMaker. It will mainly consist of the following steps:

- Data Integration on S3
- Data Exploration
- Data Cleaning and Labelling
- Feature Engineering (could be word counts, tf-idf…)
- Data Sampling for getting Training, Validation and Test Datasets to be saved again on S3
- Model development and validation with the metrics described above
- Model deployment on an endpoint to be called from a simple HTML webpage.
  In this page, one could be able to post a text to see whether is a true or fake news

Thank you for your time spent on reading my proposal.

Andrea Guidi