Udacity – Capstone Proposal

Thura Aung

Domain Background

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

Fake news has been occurring for several years; however, there is no agreed upon definition of the term "fake news". To better guide the future directions of fake news detection research, appropriate clarifications are necessary.

Problem Statement

Fake news detection is still in the early age of development, and there are still many challenging issues that need further investigations. It is necessary to discuss potential research directions that can improve fake news detection and mitigation capabilities. Binary classification is done by using different machine learning algorithms.

Datasets and inputs

I used a kaggle dataset for fake news classification purposes.

Dataset Link : 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

I first vectorize using TF-IDF vectorizer and train with machine learning algorithms such as Naive Bayes, Logistic Regression, Random Forest, Decision Tree and SVMs (Linear and RBF Kernels) for binary classification of Fake and real news.

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. The confusion matrix will also be uploaded.

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 via github repository
- Data Exploration
- Data Cleaning and Labelling
- Feature Engineering (could be 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

Thank you for your time spent on reading my proposal.

Thura Aung