ISyE 7406 - Spring 2021
Project Proposal

Team Member Name: Qi Hao, Yipei Zhang

Project Title: Detecting Fake News

1 Problem Statement

Information technology nowadays has generated an exploding amount of information. Headlines and smartphone push notifications guide us in forming opinions and making decisions in our life. However, recent events, especially the outbreak of COVID-19, unveiled the damage to our society when a piece of false information spreads. Because of the surging need to detect misinformation, researchers are studying this issue using the latest technologies in natural language processing.

Rubin et al. classified fake news into three categories: serious fabrications, large-scale hoaxes, and humorous fakes [Rubin et al., 2015]. Each of the three types poses different challenges in a successful detection at scale. Some of the later researches targeted this issue by analyzing news either under a linguistic viewpoint or using network approaches [Conroy et al., 2015]. Others developed different fake news detectors using various machine learning techniques [Gilda, 2017, Ahmed et al., 2017]. Machine learning classification algorithms that are widely studied and proven to be effective include Random Forest, SVM, Naive Bayes, Decision Trees, Stochastic Gradient Descent, and Gradient Boosting.

Our project will focus on finding a machine learning solution to the challenge of automatically detecting fake news by answering the following research questions:

- What are the indicators of fake news?
- How well can we classify fake vs. real news?

2 Data

News is the major information from which people get to know about the world. In this contemporary society, with the development of technology, every day we are involved in false or misleading information presented as news. Since tried to undertake this course project more about the field of Natural Language Processing, doing the research regarding fake news detection can be a trending topic to work on.

There are two raw data sets of our project which can be found in Kaggle [Kaggle, 2020] [1]. The first one is named Fake.csv which contains 17903 observations of fake news while the other one named True.csv is with 20826 observations of True news. Both data sets have four features:

- Title: title of the article
- Text: the content of the article
- Subject: subject related to the article
- Date: date at which the article was posted

Our goal is to use the data sets to find a classification model able to determine if an article is a fake news or not.

3 Methodology

Before proposing the candidate models for our classification task, we will first need to transform the text data into a numeric representation. We attempt to take the necessary steps, including stopword removal, lemmatizing and stemming to a root word, etc, for data preprocessing.

Our project will provide a solution for detecting fake news using a supervised learning approach. The candidate models include Logistic Regression, Naive Bayes, Linear SVM, Kernel SVM, Random Forest, Gradient Boosting, and Neural Network. We will evaluate the model performance using metrics like prediction accuracy, precision, recall, and f1 score.

[1] https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset

References

[Ahmed et al., 2017] Ahmed, H., Traore, I., and Saad, S. (2017). Detecting opinion spams and fake news using text classification. *Security and Privacy*, 1:e9.

[Conroy et al., 2015] Conroy, N. J., Rubin, V. L., and Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. In *Proceedings of the 78th ASIST Annual Meeting: Information Science with Impact: Research in and for the Community*, ASIST '15, USA. American Society for Information Science.

[Gilda, 2017] Gilda, S. (2017). Notice of violation of ieee publication principles: Evaluating machine learning algorithms for fake news detection. In *2017 IEEE 15th Student Conference on Research and Development (SCOReD)*, pages 110–115.

[Kaggle, 2020] Kaggle (2020). Fake and real news dataset — kaggle. https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset. (Accessed on 3/20/2021).

[Rubin et al., 2015] Rubin, V. L., Chen, Y., and Conroy, N. J. (2015). Deception detection for news: Three types of fakes. In *Proceedings of the 78th ASIST Annual Meeting: Information Science with Impact: Research in and for the Community*, ASIST '15, USA. American Society for Information Science.