Detecting Fake News with Supervised Learning

Benjamin R. Perucco

December 30, 2020

1 Definition

1.1 Project Overview

1.2 Problem Statement

1.3 Metrics

2 Analysis

2.1 Data Exploration

2.2 Algorithms and Techniques

2.2.1 n-gram Model

The n-gram defines usually a contiguous sequence of words with length n. For example, if $n = 1$, we speak of a unigram that contains only single word tokens. Or if $n = 2$, we denote this as a bigram that is build on two adjacent word tokens.

Consider the following text: "Sometimes we eat green apples, and sometimes, the apples we eat are red." Based on a unigram, we obtain a set of tokens: {'sometimes', 'we', 'eat', 'apples', 'green', 'and', 'the', 'are', 'red'}. We can derive a frequency array of tokens in the text: [2, 2, 2, 2, 1, 1, 1, 1, 1]. For the bigram, another set of tokens is obtained: {'sometimes we', 'we eat', 'eat green', 'green apples', 'apples and', 'and sometimes', 'sometimes the', 'the apples', 'apples we', 'eat are', 'are red'}. The corresponding frequency array of tokens in the text is: [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1].

In order to build frequency arrays for a set of texts (or documents), we need to build a common vocabulary of which the n-gram model is underlying principle.

2.2.2 Vocabulary

Consider a corpus D which contains a set of documents $\{d_1, d_2, d_3, ..., d_n\}$. Then a vocubulary F is a set of tokens $\{f_1, f_2, f_3, ..., f_m\}$ extracted from the corpus D. Please remind yourself that a token is created on the n-gram model. Usually for a set of tokens, we consider the m mostly occuring tokens in a corpus D. In the following, the tokens are denoted as features as these build the features (or independent variables) of a machine learning model.

2.2.3 Definitions

Let $\sigma(f_j, d_i)$ the number of occurances of feature f_j in document d_i. Then we can build a feature matrix

$$X_\sigma = \begin{bmatrix} \sigma(f_1, d_1) & \sigma(f_2, d_1) & \sigma(f_3, d_1) & ... & \sigma(f_m, d_1) \\ \sigma(f_1, d_2) & \sigma(f_2, d_2) & \sigma(f_3, d_2) & ... & \sigma(f_m, d_2) \\ \sigma(f_1, d_3) & \sigma(f_2, d_3) & \sigma(f_3, d_3) & ... & \sigma(f_m, d_3) \\ ... & ... & ... & ... & ... \\ \sigma(f_1, d_n) & \sigma(f_2, d_n) & \sigma(f_3, d_n) & ... & \sigma(f_m, d_n) \end{bmatrix}. \tag{1}$$

For a specific element in row i and column j (representing the document d_i and feature f_j) in the matrix X_σ, we abbreviate by using σ_{ij}.

2.2.4 Term Frequency Model

2.2.5 Inverse Document Frequency Model

2.3 Benchmark

3 Methodology

3.1 Data Preprocessing

3.2 Implementation

3.3 Refinement

4 Results

4.1 Model Evaluation and Validation

4.2 Justification

5 Conclusion

5.1 Reflection

5.2 Improvement