Machine Learning Engineer Nanodegree

Capstone Proposal: Classification of news as reliable versus unreliable

Amin Momin

June 12, 2018

Domain Background

Natural language processing (NLP) is a growing field that aims to understand textual information. The common approaches to perform NLP employ Bayesian statistical models such as naive bayes and deep learning frameworks such as recursive neural networks (RNN). While naive bayes algorithm use the probability of words to classify text information, RNN relies on complex deep learning frame work. In addition to finding word occurrence, RNN's can are trained to find similar sentences as well as relate the context of the words within the text. The use of word embedding and long short -term memory (LTSM) has greatly improved the performance of RNN in text classification algorithm.

This approach has been successfully applied to classify news articles, spam emails, evaluate yelp reviews, movie reviews as well as language translation (1,2,3,4,5).

References:

1) Text News Classification System using Naïve Bayes Classifiers http://ijoes.vidyapublications.com/paper/Vol13/39-Vol13.pdf
2) Naïve bayes and text classification https://sebastianraschka.com/Articles/2014_naive_bayes_1.html
3) Comment Abuse Classification with Deep Learning: https://web.stanford.edu/class/cs224n/reports/2762092.pdf
4) Recurrent Convolutional Neural Networks for Text Classification: https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/download/9745/9552
5) Tweet modeling with LSTM recurrent neural networks for hashtag recommendation https://ieeexplore.ieee.org/document/7727385/

Problem Statement

Fake news or unreliable news is often published for political or monitory gains by unreliable or nefarious sources (6). The proliferation of blogs, social media, and image and video sharing has exponentially increased sources of fake information. Unlike professional news organizations, these crowd sourcing or anonymous platforms disseminate information with few checks for authenticity. This has impacted hate speech, public policy, elections and safety.

Facebook and Google have been the most prominent examples of online platform often used to spread false article and headline. During the 2016 US presidential elections foreign sources spread millions of fake news articles targeting 126 million Facebook users in the United States (7). A study by Buzzfeed discovered that the fake news articles received more views than true news stories (8)

To counter this false propaganda assault on democracy the current project aims to build an algorithm to classify news articles and statements as authentic versus fake using machine learning and deep learning approaches. These methods will have the potential to understand the sentiment and context of statement and classify the category of the news articles.

Reference:

6) Fake news – Wikipedia https://en.wikipedia.org/wiki/Fake_news
7) Facebook to expose Russian fake news pages https://www.bbc.com/news/technology-42096045
8) Fake news stories make real news headlines. https://abcnews.go.com/Technology/fake-news-stories-make-real-news-headlines/story?id=43845383

Datasets and Inputs

For the present project I will use a dataset provided by a Kaggle project Fake news classification challenge (https://www.kaggle.com/c/fake-news). The training subset contain 20800 text articles and labels, while the test data contains 5200 similar information without the label. The goal of the project is to build a suitable classification model using the training data and validates the findings on the test dataset. The dataset has been manually curated to the labels.

- id: unique id for a news article
- title: the title of a news article
- author: author of the news article
- text: the text of the article; could be incomplete
- label: a label that marks the article as potentially unreliable
    - 1: unreliable
    - 0: reliable

The dataset contains relevant input information to classify the articles, such as title of the article, authors name and text of the article. Unreliable article frequently have titles with controversial text or attention grabbing keywords to get the reader's attention. Certain authors are frequently associated with unreliable sources. In addition, poorly written article have frequently associated with un-reputed news sources. Thus, having the variable of title, author and article text are valuable inputs for building news classification model using machine learning approaches.

Solution Statement

In order, to build and test the fake news classification system the training dataset will be split into 60% (12500) training, ~20% (4000) validation and ~20% (4300) test sets respectively. The model will be constructed using the training subset and validated with the validation set. The evaluation of the model will be performed using the testing subset. The accuracy of the model will be measured by comparing the true positives predictions to total number of samples use in the test dataset. The original dataset provides a test set, but it lacks labels. Therefore, we are unable to use it in the current analysis.

Benchmark Model

Naïve bayes is a valuable and efficient approach previously used for text classification. This method will be employed to determine a benchmark precision and recall for classification of news stories in the current study.

Evaluation Metrics

Once the predictions for the test article are generated using the RNN-DL model, they will be compared to the actual labels. The evaluation will be based on the total correctly classified articles compared to the total number of articles using precision (P) and recall (R) (6).

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

Additionally, we will compare results from naive Bayes with RNN-DL method for performance.

References:

9) Neural joint model for entity and relation extraction from biomedical text https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1609-9

Project Design

The project will be implemented in the following steps (pseudocode). Fig 1. is an approximate schematics representation of the steps involved in the model construction and testing

- Read the original training data and clean it
- Encode the words as into integer vectors
- Split the training set 60% (training), 20% (validation) and 20% (testing)
- Build and validate a RNN deep learning model for text classification
- Validate the model with subset of the data.
- Test the accuracy of the model prediction and compare that to original labels

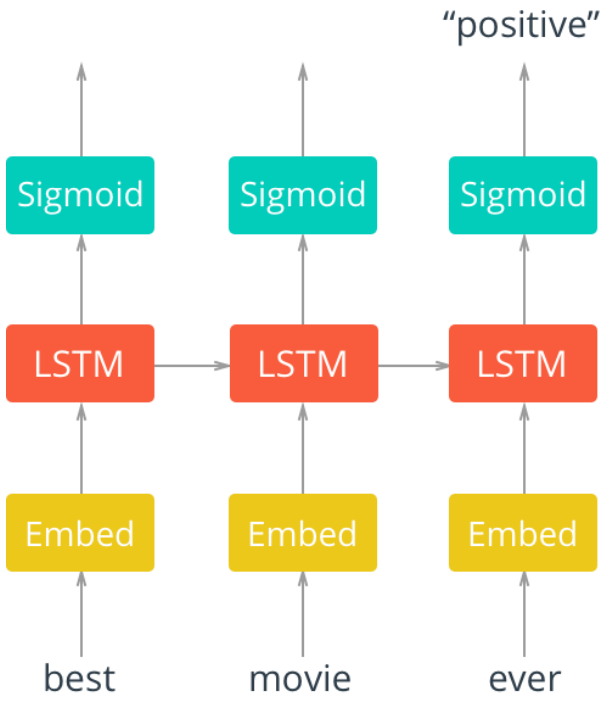

Fig 1. Schematic representation of an example RNN model that includes embedding and LSTM for sentiment analysis. Image source : https://towardsdatascience.com/sentiment-analysis-using-rnns-lstm-60871fa6aeba