Low Quality News Identification

Linfang Yang
UC Berkeley / MIDS W266 Monday
yanglinfang@berkeley.edu

Ricardo Barrera
UC Berkeley / MIDS W266 Monday
ricardofrank@berkeley.edu

Abstract

In light of recent events with low quality news, and their impact on the major political events (such as presidential election), we built a low quality news identification system where users can submit English news article text, and we would return a rating on news quality, as well a breakdown ratings on several aspects of news quality. This helps solving a particularly acute problem that needs to be addressed immediately and scalably. The exponential growth in media will further exacerbate the problem so finding a scalable way to identify news quality is vital to societal health. In this article we first explored possible areas to evaluate news quality, then we focused on selected key areas and built a Convolutional Neural Network model to automatically rate the news quality. At the end we evaluated our system using data with manually generated labels.

1 Credits

After exploring several publicly available datasets, we decided to use dataset generously shared by NOVA Search, the News Quality Dataset. Thanks to Joao Magalhaes, Ioannis Arapakis etc. who kindly helped us understanding their publication and helped us finding their dataset online. We also thanks to Denny Britz, who kindly shared his text classification approach using Convolutional Neural Network. His methods have been widely used for model benchmark, and helped us to set a decent start point of our research.

2 Introduction

News quality can be very hard to quantify, even for professional editors, but despite the difficulty, it is becoming increasing important to measure quality at scale. Recent study shows for adults in America, more than 30% news sources are from online sources. For young adults between 18-29 years old, up to 50% news sources are from online. This is very different from traditional printed newspaper. The nature of online news differs from traditional newspapers in two ways. First, scalability. It is extremely easy for an online article to reach millions of read if it is eye-catching. Second, validity. Traditional news articles often needs to be vetted by professional editors, and publishers risk their reputation if they allow fake or low quality news to be published. But online news often have diverse sources, many are hard to evaluate credibility. Even if the content of news turned out to be fake or low quality, it is hard to trace back to the original source. Evaluate news quality at scale in real-time will help solve this problem.

3 Available Datasets

There are very little datasets publicly available online for news quality. Initially we explored ways to scrap data from popular sites via a scraper / by hand then manually label the dataset. We soon realized this is not realistic and will be time consuming. Hence, we explored the following online datasets.

3.1 Kaggle Dataset

This dataset is from a kaggle `https://www.kaggle.com/mrisdal/fake-news`. It contains text and metadata from 244 websites and represents 12999 posts in total from the past 30 years, pulled using webhouse.io API. Then it used BS Detector to label the dataset as fake or not. We found that the dataset is heavily skewed towards not fake. It is understandable that there are less fake news than actual news on the web. But we prefer using dataset generated by human rather than by another labeling system. So that we do not inherit bias in the BS Detector.

3.2 Yahoo News Quality Dataset

This dataset is from NOVE Search http://novasearch.org/datasets/. It contains 500 news articles and manual labels from professional linguist or editor. Each article contains ratings on 14 aspects of news quality. Including 'Fluency', 'Conciseness', 'Descriptiveness', 'Novelty', 'Completeness', 'Referencing', 'Formality', 'Richness', 'Attractiveness', 'Technicality', 'Popularity', 'Subjectivity', 'Positive Emotion', 'Negative Emotion', 'Quality'. All ratings are in a scale of 1 to 5. All ratings are done via proper annotation and went through inner-annotator agreement, and the dataset contains score on annotator confidence as well. We decided to use this data because of the high quality of data. However, we did not use the scale 1 to 50, instead, score 1 to 3 are treated as low quality in our research. We took this approach to lower the difficulty of this project given the time constraint.

4 Data Processing and Algorithm

With well formatted dataset with labels, we started benchmarking using Kim Yoons Convolutional Neural Networks (CNN) for Sentence Classification. This is a popular benchmarking approach for text classification. We first process the dataset into input X (article full content) and input Y (two dimensional array, the encoded label for low quality or not). Then feed the dataset to the model for training. The CNN model consists of several layers. First layer embeds words in article content into low-dimensional vectors. The second layer performs convolutions over the embedded word vectors using multiple filter sizes. At the end, we max-pool the result of the convolutional layer into a long feature vector, added dropout regularization and classify the result using a softmax layer. Table 1 shows our initial results on all 14 aspects of news quality. The last row shows the overall quality classification result.

Type	Evaluation Acc	Best Train Acc
Formality	0.830769	0.84
Fluency	0.676923	0.7
Conciseness	0.630769	0.64
Descriptiveness'	0.769231	0.771739
Novelty	0.815385	0.815385
Completeness	0.676923	0.77
Referencing	0.8	0.8
Richness	0.846154	0.846154
Attractiveness	0.892308	0.91
Technicality	0.861538	0.89
Popularity	0.861538	0.87
Subjectivity	0.923077	0.99
Positive Emotion	0.969231	0.99
Negative Emotion	0.969231	0.969231
Quality	0.846154	0.846154

Table 1: News quality classification results on different aspects.

5 Next Steps

After setting up benchmarks, we realized that some language aspects are easier to classify than others. For example for subjectivity, model prediction reached 0.92 accuracy. While for conciseness, it only reached 0.63 accuracy. We will explore whether all 14 aspects of languages contribute to final news quality. Or just some aspects. After locking on which aspect we mainly focus on improving, we will work on improving the model, potentially try other neural network models.

Acknowledgments

Thanks again to Joao Magalhaes, Ioannis Arapakis, Denny Britz who helped us getting datasets, and setting benchmarks.

References

https://www.researchgate.net/profile/Victoria_Rubin/publication/281818851_Deception_Detection_for_News_Three_Types_of_Fakes/links/55f96a9b08aeafc8ac24260e.pdf

https://pdfs.semanticscholar.org/b112/1655f0858ba38226a4dd614e84cb86c2a8a6.pdf

http://ethesis.nitrkl.ac.in/3578/1/thesis.pdf

http://www.aclweb.org/anthology/W/W16/W16-0802.pdf

http://www.journalism.org/2016/07/07/pathways-to-news/

https://www.kaggle.com/mrisdal/fake-news

http://novasearch.org/datasets/

http://www.wildml.com/2015/12/implementing-a-cnn-for-text-classification-i