# Scalable 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

Convolutional Neural Networks (CNNs) greatly improve modeling sophistication and power which enables certain scenarios in Natural Language Processing such as document classification and feature extraction to perform adequately for consumer use. In this paper, we discuss leveraging existing work using CNNs to rate articles along fourteen dimensions of labels describing article quality (e.g. 'Conciseness') and explore the possible connection between article veracity and quality via correlations. The motivation for this effort stems from the exponential growth in media that will further exacerbate the problem of misinformation overload.

## 1 Introduction and Background

Modern society suffers from an overwhelming amount of information due to technological advancements that make incessant and ubiquitous media consumption the norm. As a result, consumers now have a uniquely modern challenge to filter and validate their information sources to avoid misleading information, which we call 'Low Quality News'.

Unfortunately, it is unreasonable to expect individuals to filter and verify all of the incoming information. This presents and opportunity to assist individuals by automatically labeling content to quickly and easily to identify low quality information, similar to but not entirely described by the colloquial term, 'Fake News'.

This project leverages existing work to classify articles via Convolutional Neural Networks (CNNs) based upon fourteen different dimensions (e.g. 'Comprehensiveness' and 'Conciseness') in an ensemble fashion to identify low-quality news articles. *[6]* The long-term vision is to eventually create a news categorization platform that will allow people to filter and curate their content quickly, easily, and reliably.

The evaluation and exploration process involved gathering many Facebook articles that were already labeled as a part of a BuzzFeed article dataset examining the quality of major information sources on Facebook, with the goal of being able to identify correlations between article quality and article veracity, identified with the categories: 'Mostly True', 'Mix of True and False', 'Mostly False', and 'No Factual Content'. *[5]*

## 2 Datasets Selection Process

There are very few datasets publicly available online with labeling for article quality and veracity, and generating new datasets is expensive and time consuming. Therefore, we worked with whatever datasets we could find. If the effort showed promise, it would then prove worthwhile to fund and generate a new dataset in-house and further investigate.

### 2.1 Fake News Kaggle Dataset

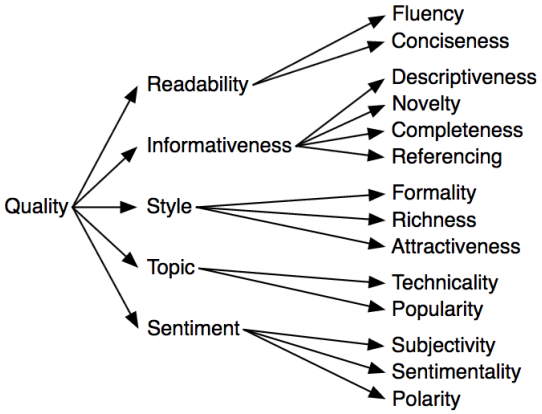
https://www.kaggle.com/mrisdal/fake-news

We did not end up using this dataset in our final results for various reasons, but it is worth exploring later since it provides automatically labeled data, which can be useful to evaluate a more mature system later on. 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. We preferred using dataset generated by human rather than by another labeling system so that we did not inherit bias in the BS Detector.

### 2.2 Yahoo News Quality Dataset

`http://novasearch.org/datasets/`

This dataset is from NOVE Search . It contains 500 news articles and manual labels from professional linguist or editor. Each article contains ratings on 14 aspects of news 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 5, 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.




### 2.3 BuzzFeed Article Veracity Dataset

`https://github.com/BuzzFeedNews/2016-10-facebook-fact-check`

This dataset is from BuzzFeed and contains links to a few thousand Facebook news articles and has manual labels done by BuzzFeed staff. Each article is labeled as 'Mostly True', 'Mix of True and False', 'Mostly False', and 'No Factual Content'. We decided to use this data because it was the only dataset we could find regarding article veracity.

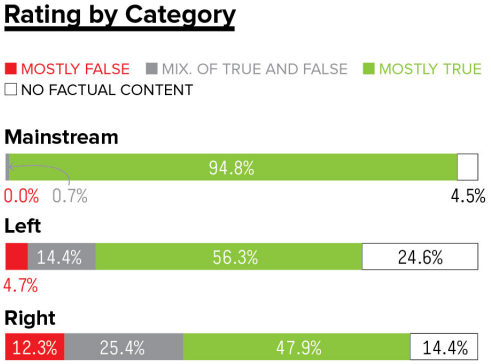
## 3 Background & Related Work

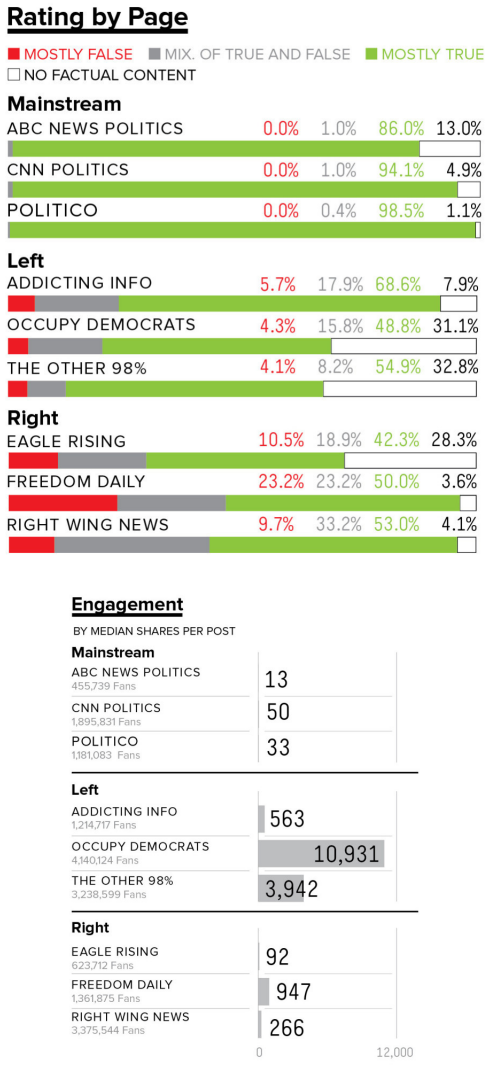
Interest in identifying and fighting 'Fake News' is surging in 2017 as a result of an ever-worsening polarization of discourse and media in modern society. The effort to fight misinformation is being led by some noteworthy groups such as Facebook, the US Govt, the EU, & NATO. [1] [2] [3] [4] Their approaches vary since there is no clear right solution, but technology powered by Artificial Intelligence (AI) is a requirement for every solution / effort.

Our project is focused on fighting misinformation and we started by re-purposing similar work done by Arapakis, Peleja, Cambazoglu, & Magalhaes. Their goal involved managing article quality by training a CNN to label an article through fourteen dimensions as a 'gatekeeper' standard for articles to be allowed to go online.

Their approach, however, is not as effective as it could be because it assumes that writers are incompetent and editors are trying to maintain high article quality. Our group believes this project is best approached like a security problem whether the writer is analogous to a malicious attacker, where a model or system would be used defensively by the consumer to protect against malicious news feeds.

This assumption appears to have merit as BuzzFeed conducted an extensive study regarding articles published by highly partisan Facebook pages (e.g. 'Eagle Rising' and 'Occupy Democrats'). The results indicate that partisan pages are much more widely shared and had significant less veracity than mainstream articles.






Therein lies the crux of the problem, and our project aims to alleviate the scope and impact of misinformation campaigns.

## 4 Methods and Results

We established a benchmark using Kim Yoons 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, one-hot encoded for article quality) and 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.

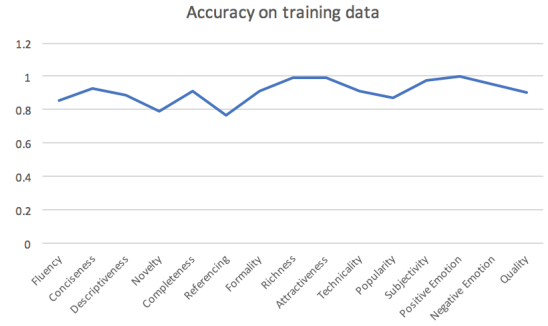
After benchmarking, we trained fourteen separate CNNs with an 80/20 train-dev split with 6-fold validation. The purpose was to look for correlations between the article quality features and article veracity on the BuzzFeed article veracity dataset. By using the trained quality dimension models to predict the BuzzFeed article quality attributes, we could then see if certain attributes were more present in truthful articles as opposed to untruthful ones.

The process involved manually extracting article text from the BuzzFeed Facebook article links with somewhat-balanced representation for all of the categories. We ended up with 159 articles ranging from 'Mostly True' to 'Mostly False' that were hand labeled by BuzzFeed's dataset creators.

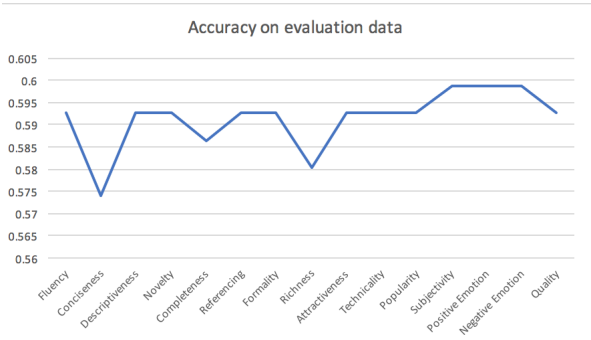
We then passed the BuzzFeed articles to each of the trained CNNs to see what the model accuracy / correlation was when predicting veracity given a model trained on a specific dimension of quality.

## 5 Results

The results on the training and test data look quite good for predicting the fourteen different dimensions:




The results seem poor, however, when extending the Article Quality models to the Facebook BuzzFeed dataset to look for correlation between a quality dimension and the veracity of an article. It appears fair to say, however, that the results are inconclusive at the moment due to lack of data and because other options exist to possibly make a working and useful model to connect quality and veracity.




## 6 Conclusion

It is definitely possible to label articles accurately based upon the fourteen dimensions of quality discussed in this paper, and there may be some merit in using article quality in conjunction with the article content to automatically judge article veracity. This evidence is mostly anecdotal, however, based upon our experience sifting through the BuzzFeed Facebook articles–there was a clear writing style and choice of diction associated with untruthful articles.

Unfortunately, we ran out of time and did not get to a working ensemble model to accurately judge veracity as we envisioned, but we believe further work on this can and will bear fruit. The next best course of action is to properly create and curate sufficient high-quality content to evaluate the connection between quality and veracity with the working Article Quality model and then build a working ensemble that consumes the article content (e.g. text, website, etc) along with the set of quality features (e.g. conciseness) to create an accurate predict on veracity ('Mostly True' or 'Not Mostly True').

## Acknowledgments

Thanks to Joao Magalhaes, Ioannis Arapakis etc. who kindly helped us understanding their publication and helped us finding their dataset online. 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.

## References

1. http://fortune.com/2017/04/05/ebay-pierre-omidyar-100m-fight-fake-news
2. http://bigstory.ap.org/article/d5b1763ae5ad463ba8bc4fff5fa23816/9-eu-nato-nations-set-center-fight-hybri
3. http://www.snopes.com/obama-signs-christmas-bill-making-alterna
4. http://www.cnbc.com/2017/04/27/facebook-to-fight-fake-news-groups.html
5. https://github.com/BuzzFeedNews/2016-10-facebook-fact-check
6. http://aclweb.org/anthology/P/P16/P16-1178.pdf
7. https://www.researchgate.net/profile/Victoria_Rubin/publication/281818851_Deception_Detection_for_News_Three_Types_of_Fakes/links/55f96a9b08aeafc8ac24260e.pdf
8. https://pdfs.semanticscholar.org/b112/1655f0858ba38226a4dd614e84cb86c2a8a6.pdf
9. http://ethesis.nitrkl.ac.in/3578/1/thesis.pdf
10. http://www.aclweb.org/anthology/W/W16/W16-0802.pdf

11. http://www.journalism.org/2016/07/07/pathways-to-news/
12. https://www.kaggle.com/mrisdal/fake-news
13. http://novasearch.org/datasets/
14. http://www.wildml.com/2015/12/implementing-a-cnn-for-text-classification-in-tensorflow/