**Fake News Detection with**

**Unsupervised Learning**

DSCC 440 Data Mining

Jordan Pappas, Isaac Manly

University of Rochester, Rochester, NY

**Abstract**

After the 2016 US presidential election in which outsider Donald Trump defeated career politician Hillary Clinton, the phenomenon of “fake news” came to the forefront of American media. More specifically, American media was accused of broadcasting fake news to impact voter behavior. This issue was exacerbated by widespread social media use, in which many users did not take the time to fact-check the information consumed from their feeds. Moreover, fake news existed within a larger ecosystem of disinformation, which degraded the common ground truth all Americans once had in common. This issue has recently reemerged due to the novel coronavirus. Producers of disinformation threaten public efforts by falsely equivocating the coronavirus with the flu, rejecting the efficacy of mask-wearing, or even denying the pandemic’s existence. As such, different methods to detect fake news are paramount. To this end, we employ unsupervised models. While supervised models have already been shown to reliably classify information veracity, we seek a novel and promising approach in various feature engineering and clustering techniques to detect fake news articles. The differing rhetoric, diction, and syntax used between accurate and inaccurate news articles allows us to leverage these linguistic discrepancies via unsupervised models.

**Introduction**

The goal of this paper is to explore what features most effectively predict article veracity and investigate what unsupervised models best predict article veracity. After the 2016 presidential election and the coronavirus pandemic, there is a newfound urgency to distinguish the truth value of news. This phenomenon’s importance is exemplified by the Oxford dictionary’s addition of “Fake News” to its catalog:

false news stories, often of a sensational nature, created to be widely shared online for the purpose of generating ad revenue via web traffic or discrediting a public figure, political movement, company, etc.

The importance of detecting fake news grows annually. This is exemplified by an increasing trend in the number of Americans who rely on social media for news. In 2019, the Pew Research Center found that 4 in 10 Americans consume news on Facebook and that 52% of survey respondents have changed their social media behavior due to the issue of fake news1. This digital growth phenomenon is two fold: Not only are news consumers increasingly relying on social media news, but also digital news sources are rapidly growing. Employment in digital newsrooms grew 82% between 2008 and 2018, while traditional newsroom employment dropped about 51% in the same time period1.

Studies have also shown that the emergence of fake news is tied to the emergence of social media. In a 2018 study, Nelson and Taneja found that the average number of monthly visitors for a fake news site was approximately 650,000 during November 2016. Although the average time spent on fake news sites was less than half the average time spent on real new sites, the fake news audience spent more overall time on the internet, especially on Facebook. Thus, the authors posit that the increasing adoption of social media is a cause for the increase in traffic for fake news sources4.

Finally, the importance of fake news detection has escalated since the coronavirus pandemic’s spread in March 2016. Ten months into the pandemic, the United States alone has experienced 280,000 deaths, a number purported by the CDC to be underestimated by 6 to 24 times that amount. As such, mask mandates and social distancing guidelines have been implemented. However, some citizens, many of whom have been influenced by fake news, have taken to resisting these enforcements.The adoption and spread of fake news may be compelling to these audiences because of human tendencies such as confirmation bias6. The emergence of fake news regarding COVID-19 has been found in Italy. Early in the pandemic, Italy was one of the hardest hit countries. A group of researchers at the University of Florence studied the circulation of fake news during a 5 month period in 2020. Of their sample, 23.1% of website links shared directed to a website that was considered fake. They found that fake news “indubitably” impacted health communications during a pandemic6. In a global health crisis, implementations of fake news detection models by social media providers may save lives.

**Related Works**

Fake news detection methods are predominantly implemented using supervised learning techniques. In 2019, Khan, Afroz, et al. studied the efficacy of popular classification methods. Their models focused on the content of the news articles which includes the word count, subject, sentiment, number of nouns, etc. of the text within the articles. Their accuracies ranged from 0.54 to 0.95, with Naive Bayes performing the best3. In our paper, we will differ from Khan’s approach by using unsupervised instead of supervised learning and exploring the effect of context, not just article content.

Fake news detection methods have also been implemented using network graph models. A study by Alexandre Bovet and Hernan Makse used this approach to determine that during the 5 months preceding the 2016 election, 25% of news-related tweets spread either fake or extremely biased news. They found significant differences between graph densities given political leaning. Moreover, they discovered that conservative networks have higher homogenous connectivity than their counterparts2. This finding is useful since, like Bovet, we will use the news article context to explore authenticity, ie. political leanings may be useful for clustering the data.

Fake news detection methods have also been implemented using Bayesian network models. In 2019, Yang, Shu, et al. leveraged a Bayesian network model to capture the conditional probabilities between the truth of news, the user’s opinion, and the credibility of news. Their method used both the context and content of tweets to model the truth of user’s posts. Their accuracy was 0.76 on the Liar dataset, and 0.68 on the Buzzfeed dataset7.

Another group used a graph-based approach in 2020. Long, Gangireddy, et al. combined biclique identification, graph-based feature vector learning and label spreading to achieve an accuracy of 0.77 on the GossipCop dataset8.

**Methodology**

Traditionally, supervised learning methods have been used to tackle fake news detection, but there is one serious limitation to this family of techniques: the data must be reliably labeled before training. Unsupervised learning methods, however, avoid this limitation, as they don’t require labels. Additionally, unsupervised learning will be more generalizable than supervised learning, because supervised learning will only perform well on news topics that it has been trained on beforehand, while unsupervised learning requires no training at all. Accordingly, we will explore the predictive value of unsupervised clustering, including hierarchical and partitioning methods, using both the content and context of news article data.

For our modeling, we used two datasets: Liar9 and Kaggle Fake News10. Liar contains approximately 13,000 short news statements and labels collected by Politifact.com. The Liar dataset is useful because it contains many context features that are impossible to ascertain from the data such as political affiliation, speaker, location, etc. The Kaggle Fake News dataset is from a Kaggle competition and it has approximately 40,000 news articles with their headline, body, subject and date. This dataset is useful for extracting content because of its size and variety of text. In order to understand the subjects covered in the data, we performed LDA topic modeling:

| Kaggle | {house, white, news}  {president, obama, american}  {government, united, reuters}  {trump, president, donald}  {election, republican, campaign} |
| --- | --- |
| Liar | {jobs, million, americans}  {texas, united, voted}  {federal, government, law}  {obama, health, president}  {tax, billion, taxes} |

Figure 1.

Topic modeling shows that our data is predominantly political in nature.

In order to engineer features on the content of the articles, we used linguistic inquiry and word count (LIWC), sentiment analysis, and clustering of term frequency - inverse document frequency (tf-idf). For LIWC, we implemented an NLP package named spaCy. Figure 2 is an example of how spaCy processes language.

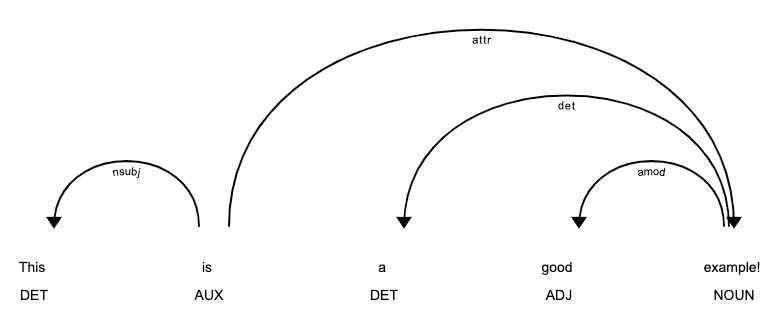


Figure 2.

For each observation, the number of adverbs, nouns, adjectives, and proper nouns out of the total word count were calculated.

To capture the sentiment of a text article, Vader Sentiment Analysis was implemented which created a number of features such as a sentiment score from -1 to 1, 1 being the most positive score. Vader sentiment analysis relies on a dictionary that maps lexical features to emotion intensities.

The clustering of tf-idf involved applying KMeans clustering to word frequency vectors of the text. The number of clusters (K) was chosen with the Elbow Method seen in Fig. 3.

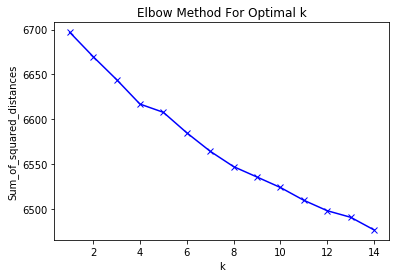


Figure 3.

This feature is intended to group similar texts together, using their cluster label as the feature. There was no clear elbow, so we chose K = 8. Once clustered, we used PCA to transform the clustered data into 2 components for graphing in Figure 4.

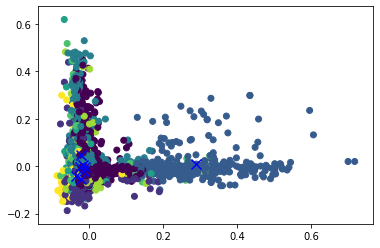


Figure 4.

The Liar dataset was chosen because it was produced with context included. The context provided was the political party of the speaker, the speaker, the topic, the location, and the position of the speaker. The kaggle fake news dataset was chosen because of how many news samples it contained and their quality.

**Experiments**

One of the methods explored is BIRCH. The features used will include content, context, and combination of content and context from the Liar dataset. Next, the algorithm will be run on only the content features of the Kaggle dataset, as it has no context features. This procedure will be implemented for KMeans as well with a K = 2 as we are implementing the algorithm as a binary classifier. BIRCH is implemented in this manner as well. The pseudocode for KMeans is in the figure below:

| **Input:**  *k*: the number of clusters,  *D*: a data set containing *n* objects.  **Output:** A set of *k* clusters.  **Method:**  1. arbitrarily choose *k* objects from *D* as the initial cluster centers;  **2. repeat**  3. (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;  4. update the cluster means, that is, calculate the mean value of the objects for each cluster;  **5. until** no change; |
| --- |

Figure 6. From the Morgan Kauffman textbook11

The evaluation for these models will be similar to that of standard classifiers. Since we are attempting to differentiate our data into two clusters, true and false, those cluster labels will be interpreted as their class label. For instance, a perfect unsupervised method, consisting of only true articles in one cluster, and false in the other, would have an accuracy of 1.

**Results**

**Context**

The results for KMeans and BIRCH clustering are found in figure 7.

| Liar | Accuracy | Precision | Recall |
| --- | --- | --- | --- |
| BIRCH | 0.58 | 0.74 | 0.63 |
| KMeans | **0.59** | **0.74** | **0.65** |

Figure 7.

The two methods perform similarly on Liar, but partitioning performs slightly better. The content features used included party, word count, location of speaker, and the topic of the speaker.

**Content**

The results for KMeans and BIRCH clustering are found in figure 8. Both methods perform decently on the kaggle dataset with respect to accuracy. However, BIRCH performance drops rapidly on the Liar dataset. This is due to BIRCH assigning the majority of observations to one cluster. KMeans performs well on the content of the Liar dataset with an accuracy of 0.69.

| Kaggle | Accuracy | Precision | Recall |
| --- | --- | --- | --- |
| BIRCH | 0.61 | 0.03 | 0.00 |
| KMeans | 0.59 | 0.22 | 0.02 |
| Liar | Accuracy | Precision | Recall |
| BIRCH | 0.26 | 0.00 | 0.00 |
| KMeans | **0.69** | **0.75** | **0.89** |

Figure 8.

The content features used were the percentage of different parts of grammar, word count, sentiment, and tf-idf cluster.

**Content and Context**

The performances, shown in figure 9, of BIRCH and KMeans are poor on the combination of content and context features.

| Liar | Accuracy | Precision | Recall |
| --- | --- | --- | --- |
| BIRCH | 0.45 | **0.77** | 0.37 |
| KMeans | **0.54** | 0.72 | **0.61** |

Figure 9.

This is likely due to the high dimensionality of combining those features and the added noise from each additional feature. The features used were the features from both the content and context sections above.

**Conclusion**

We approached the task of fake news detection with unsupervised learning by using their content and context. This task is considerably more difficult than implementing supervised learning. To engineer our content features, we employed sentiment analysis, clustering of tf-idf’s, and LIWC from two datasets: Liar and Kaggle Fake News. After extracting new content features, we tested algorithms BIRCH and KMeans performance on the context, the content, and then both combined.

The best results were achieved by KMeans using the content of the Liar data, resulting in an accuracy of 0.69, a precision of 0.75, and a recall of 0.89. For this task, partitioning methods are likely better than hierarchical clustering.

Future work in the area should explore alternative methods of feature extraction from the content of text. For example, we did not represent all parts of speech from the text. Other types of unsupervised learning should also be tested, especially those that cannot be implemented as a binary classifier.

**Bibliography**

[10] Bisaillon, Clément. “Fake and Real News Dataset.” *Kaggle*, 26 Mar. 2020, www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset.

[2] Bovet, Alexandre, and Hernán A. Makse. “Influence of Fake News in Twitter during the 2016 US Presidential Election.” *Nature News*, Nature Publishing Group, 2 Jan. 2019, www.nature.com/articles/s41467-018-07761-2.

[5] Darley, John, and Paget Gross. “A Hypothesis-Confirming Bias in Labeling Effects.” *Dimensions*, 1983, app.dimensions.ai/details/publication/pub.1001508409.

[1] Geiger, A.W. “Key Findings about the Online News Landscape in America.” *Pew Research Center*, Pew Research Center, 30 May 2020, www.pewresearch.org/fact-tank/2019/09/11/key-findings-about-the-online-news-landscape-in-america/.

[11] Han, Jiawei., Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques, Third Edition*. 3rd ed. Waltham, Mass.: Morgan Kaufmann Publishers, 2012.

[3] Khan, Junaed Younus, et al. “[PDF] A Benchmark Study on Machine Learning Methods for Fake News Detection: Semantic Scholar.” *arXiv*, 1 Jan. 1970, www.semanticscholar.org/paper/A-Benchmark-Study-on-Machine-Learning-Methods-for-Khan-Khondaker/09adaadcd4cc3e7aafef6573df7cab133cb9517b.

[6] Moscadelli, Andrea, et al. “Fake News and Covid-19 in Italy: Results of a Quantitative Observational Study.” *MDPI*, Multidisciplinary Digital Publishing Institute, 12 Aug. 2020, www.mdpi.com/1660-4601/17/16/5850.

[4] Nelson, Jacob, and Harsh Taneja. “The Small, Disloyal Fake News Audience: The Role of Audience Availability in Fake News Consumption.” *SSRN*, 30 Jan. 2018, papers.ssrn.com/sol3/papers.cfm?abstract\_id=3107731.

[8] Siva Charan Reddy Gangireddy Indraprastha Institute of Information Technology, et al. “Unsupervised Fake News Detection: A Graph-Based Approach.” *Unsupervised Fake News Detection | Proceedings of the 31st ACM Conference on Hypertext and Social Media*, 1 July 2020, dl.acm.org/doi/10.1145/3372923.3404783.

[9] thiagorainmaker77. “thiagorainmaker77/liar\_dataset.” *GitHub*, github.com/thiagorainmaker77/liar\_dataset.

[7] Yang, Shuo, et al. “Unsupervised Fake News Detection on Social Media: A Generative Approach.” *Proceedings of the AAAI Conference on Artificial Intelligence*, ojs.aaai.org//index.php/AAAI/article/view/4508. .