# BREAKING! Presenting Fake News Corpus For Automated Fact Checking

**Archita Pathak**
University at Buffalo (SUNY)
New York, USA
architap@buffalo.edu

**Rohini K. Srihari**
University at Buffalo (SUNY)
New York, USA
rohini@buffalo.edu

## Abstract

Popular fake news articles spread faster than mainstream articles on the same topic which renders manual fact checking inefficient. At the same time, creating tools for automatic detection is as challenging due to lack of dataset containing articles which present fake or manipulated stories as compelling facts. In this paper, we introduce manually verified corpus of compelling fake and questionable news articles on the USA politics, containing around 700 articles from Aug-Nov, 2016. We present various analyses on this corpus and finally implement classification model based on linguistic features. This work is still in progress as we plan to extend the dataset in the future and use it for our approach towards automated fake news detection[1].

## 1 Introduction

Fake news is a widespread menace which has resulted into protests and violence around the globe. A study published by Vosoughi et al. of MIT states that falsehood diffuses significantly farther, faster, deeper and more broadly than the truth in all categories of information. They also stated that effects were more pronounced for false political news than for false news about terrorism, natural disaster, science, urban legends or financial information. According to Pew Research Center's survey [2] of 2017, 64% of US adults said to have great deal of confusion about the facts in the current events. Major events around the globe saw sudden jump in deceitful stories on internet during sensitive events because of which social media organizations and government institutions have scrambled together to tackle this problem as soon as possible. However, fake news detection has its own challenges.

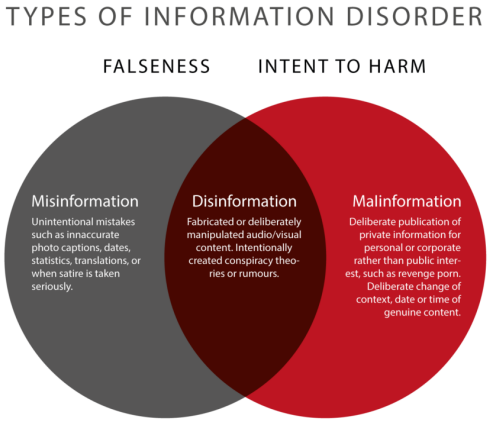


Figure 1: Conceptual framework for examining information disorder presented by Wardle and Derakhshan

First and foremost is the consensus on the definition of fake news which is a topic of discussion in several countries. Considering the complexity of defining fake news, Wardle and Derakhshan instead created a conceptual framework for examining information disorder as shown in Figure - 1. Based on this model, EU commission in 2018 categorized fake news as disinformation with the characteristics being *verifiably false or misleading information that is created, presented and disseminated for economic gain or to intentionally deceive the public, and in any event to cause public harm.* Since our work focuses on false information written intentionally to deceive people in order to cause social unrest, we will be following the definition defined by EU to categorize fake articles into various categories as defined in Section 3 of this paper.

The second challenge is the lack of structured and clean dataset which is a bottleneck for fake opinion mining, automated fact checking and creating computationally-intensive learning models

[1]Dataset is available at: https://github.com/architapathak/FakeNewsCorpus

[2]http://www.journalism.org/2016/12/15/many-americans-believe-fake-news-is-sowing-confusion/

for feature learning. In the following section, we elaborate on this challenge more, specify related works and how our work is different from others.

## 2 Related Works

There have been several datasets released previously for fake news, most notably Buzzfeed[3] and Stanford (Allcott and Gentzkow, 2017) datasets containing list of popular fake news articles from 2016. However, these datasets only contain webpage links of these articles and most of them don't exist anymore. Fake news challenge, 2017[4] released a fake news dataset for stance detection, i.e, identifying whether a particular news headline represents the content of news article, but almost 80% of the dataset does not meet the definition that we are following in our work. Many of the articles were satire or personal opinions which cannot be flagged as *intentionally deceiving the public*. (Wang, 2017) released a benchmark dataset containing manually labelled 12, 836 short statements, however it contains short statements by famous personalities and not malicious false stories about certain events that we are interested in. Finally, the dataset created by using BS detector[5] contains articles annotated by news veracity detector tool and hence, cannot be trusted, as the labels are not verified manually and there are many anomalies like movie or food reviews being flagged as fake.

In this paper, we overcome these issues by manually selecting popular fake and questionable stories about US politics during Aug-Nov, 2016. The corpus has been designed in such a way that the writing style matches mainstream articles. In the following sections, we define our motivation for creating such corpus, provide details on it and present a classification model trained on this corpus to segregate articles based on writing style. We also discuss properties of fake news articles that we observed while developing this corpus.

Other notable works like FEVER dataset (Thorne et al., 2018), TwoWingOS (Yin and Roth, 2018) etc. are focused on claim extraction and verification which is currently out of scope of this paper. In this work, we are solely focused on creating a clean corpus for fake news articles and performing classification of articles into "questionable" and "mainstream" based on writing style. Fact checking, fake news opinion mining and fake claim extraction and verification fall under future work of this paper. For classification task, previous works include Potthast et al. who used stylometric approach on Buzzfeed dataset and achieved an F1 score of 0.41; Pérez-Rosas et al. who used Amazon Mechanical Turk to manually generate fake news based on real news content and achieved an F1 score of 0.76 to detect fake news, and Ruchansky et al. who used Twitter and Weibo datasets and achieved F1 scores of 0.89 and 0.95 on respective datasets. In our work, we were able to achieve an F1 score of 0.97 on the corpus we have created, hence setting benchmark for the writing-style based classification task on this corpus.

## 3 Motivation

Fake news detection is a complicated topic fraught with many difficulties such as freedom of expression, confirmation bias and different types of dissemination techniques. In addition to these, there are three more ambiguities that one needs to overcome to create an unbiased automated fake news detector - 1. harmless teenagers writing false stories for monetary gains[6]; 2. AI tools generating believable articles[7] and; 3. mainstream channels publishing unverifiable stories[8]. Considering these elements, our motivation for this problem is to design a system with the focus on understanding the inference of the assertions, automatically fact check and explain the users why a particular article was tagged as questionable by the system. In order to do so, our first priority was to create an efficient corpus for this task. As we will see in the following sections, we have manually selected the articles that contain malicious assertions about a particular topic (2016 US elections), written with an intent of inciting hatred towards a particular entity of the society. The uniqueness of the articles in this corpus lies in the fact that a reader might believe them if not specifically informed about them being fake, hence confusing them whether the article was written by mainstream media or fake news

[3] https://www.buzzfeednews.com/article/craigsilverman/these-are-50-of-the-biggest-fake-news-hits-on-facebook-in

[4] http://www.fakenewschallenge.org/

[5] https://www.kaggle.com/mrisdal/fake-news/data/

[6] https://www.wired.com/2017/02/veles-macedonia-fake-news/

[7] https://www.technologyreview.com/s/612960/an-ai-tool-auto-generates-fake-news-bogus-tweets-and-plenty-of-gibberish/

[8] https://indianexpress.com/article/cities/delhi/retract-reports-that-claim-najeeb-is-isis-sympathiser-delhi-hc-to-media-5396414/

channel. This uniqueness will ultimately help in creating a detector which is not biased towards mainstream news articles on the same topic (2016 elections in this case.)

## 4 Corpus

### 4.1 Creation and Verification

The final corpus consists of 26 known fake articles from Stanford's dataset along with 679 questionable articles taken from BS detector's kaggle dataset. All the articles are in English and talk about 2016 US presidential elections. We, first, went through the Stanford's list of fake articles links and found the webpages that still exist. We then picked the first 50-70 assertive sentences from these articles. After that, we read the articles from BS detector's kaggle dataset and selected those articles which are written in a very compelling way and succeeded in making us believe its information. We then manually verified the claims made by these articles by doing simple Google searches and categorized them as shown in Table - 1. Based on the type of assertions, we have come up with two types of labels:

**Primary Label**: based on the assertions they make, articles are divided into 3 categories: 1. False but compelling (innovated lies having journalistic style of writing); 2. Half baked information i.e, partial truth (manipulating true events to suit agenda); and 3. Opinions/commentary presented as facts (written in a third person narrative with no disclaimer of the story being a personal opinion).

**Secondary Label**: is of 2 types: 1. articles extracted from Stanford's dataset have been tagged as *fake* as per their own description and; 2. articles taken from kaggle dataset are tagged as *questionable*. We believe tagging something fake, when they do contain some elements of truth, will be an extreme measure and we leave this task to the experts.

Finally, the corpus was cleaned by removing articles with first person narrative; removing images and video links and keeping only textual content; removing tweets, hyperlinks and other gibberish like *Read additional information here, [MORE], [CLICK HERE]* etc.

### 4.2 Analysis

**Features**: We used NLTK package to explore basic features of the content of news articles in terms of sentence count, word count etc. Number of assertions ranges from 2 to 124 with word count varying from 74-1430. Maximum number of stop words in the longest article is 707. Since keywords form an important representation of the dataset, we extracted top 20 most common keywords from the corpus as shown in Figure - 2.

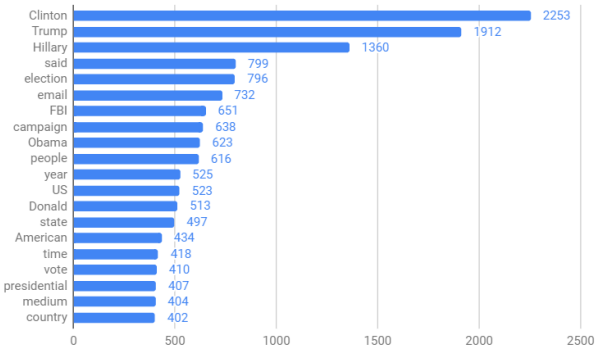


Figure 2: Top 20 most common keywords

**Comparison with mainstream article**: Mainstream articles from publishers like New York Times were extracted from *all the news*[9] kaggle dataset, which contains news articles from 15 US mainstream publishers. We selected articles from Oct - Nov, 2016 covering US politics during the time of election. There were total 6679 articles retrieved from this dataset which were then categorized into two labels as per our corpus's schema. Table - 2 compares characteristics of top 3 sentences from mainstream news articles with our corpus. We can observe that there are not many dissimilarities except the information presented in both the articles.

**Observed Characteristics**: Although articles in our corpus have many similarities with mainstream articles, there are some underlying patterns that can be noticed by reading all of them together at once. Following are the observations that we made while verifying the claims presented in these articles.

1. All of the news articles were trying to create sympathy for a particular entity by manipulating real stories. They were either trying to sympathize with Donald Trump by mentioning media rhetoric against him, or with Hillary Clinton by mentioning Trump's past.

2. Similarly, they also made false claims against above mentioned entities, by referring leaked

[9] https://www.kaggle.com/snapcrack/all-the-news

| URL | Authors | Content | Headline | Primary Label | Secondary Label |
|---|---|---|---|---|---|
| URL of article | Can contain anonymous writers | Collection of assertions | Headline of the article | 1. False<br>2. Partial truth<br>3. Opinions | 1. Fake<br>2. Questionable |

Table 1: Corpus schema. No cleaning has been performed on headlines as they are intentionally made catchy for attention seeking.

| Attributes | Mainstream | Questionable |
|---|---|---|
| Word count range (min to max) | 20-100 | 21-100 |
| Character count range (min to max) | 89-700 | 109-691 |
| Uppercase words count range (min to max) | 0-14 | 0-8 |
| Mainstream Example | WASHINGTON - An exhausted Iraqi Army faces daunting obstacles on the battlefield that will most likely delay for months a major offensive on the Islamic State stronghold of Mosul, American and allied officials say. The delay is expected despite American efforts to keep Iraqs creaky war machine on track. Although President Obama vowed to end the United States role in the war in Iraq, in the last two years the American military has increasingly provided logistics to prop up the Iraqi military, which has struggled to move basics like food, water and ammunition to its troops. | |
| Questionable Example | WASHINGTON - Hillary Clinton is being accused of knowingly allowing American weapons into the hands of ISIS terrorists. Weapons that Hillary Clinton sent off to Qatar ostensibly designed to give to the rebels in Libya eventually made their way to Syria to assist the overthrow of the Assad regime. The folks fighting against Assad were ISIS and al-Qaeda jihadists. | |

Table 2: Comparing basic features of top 3 sentences from mainstream articles and questionable articles.

documents from WikiLeaks or other similar sources. In our verification, we found that in most of these articles, only few claims matched the leaked story and rest were invented.

3. Articles first tell false stories against a certain entity and then asks the reader questions such as "Do you think mainstream media is conspiring against you by hiding this story?", "Do you think Google is making algorithms to create an illusion of reality?" After asking such questions, they either leave the reader hanging in contemplation or ask them to "share [the article] if you believe so."

4. The above point then leads to what (Nyhan and Reifler, 2010) describes as *backfire effect* in which correcting readers actually makes them believe false stories.

5. Fake news writers have learnt/are learning to use mainstream writing style such as, mentioning city name before starting the article, mentioning abbreviations like (SC) for Security Council, (AI) for Amnesty Investor, using elaborate vocabulary etc.

| Split | Random | | K-Fold | |
|---|---|---|---|---|
| | Questionable (1) | Mainstream (0) | Questionable (1) | Mainstream (0) |
| *Train* | 406 | 5334 | 396 | 5343 |
| *Test* | 90 | 1345 | 100 | 1336 |

Table 3: To meet the real world scenario, train data and test data have been split with an approximate ratio of 1:10.

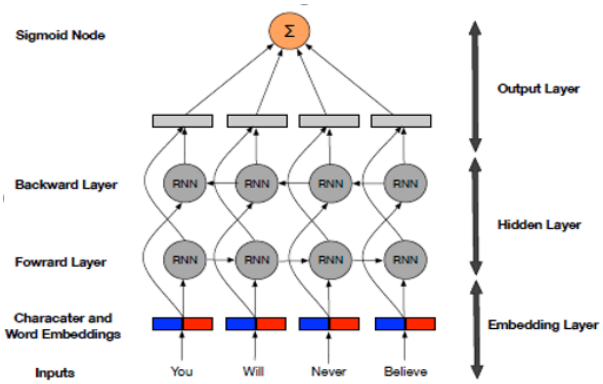


Figure 3: Bi-directional architecture

## 5 Classification Task

### 5.1 Model

After creating this corpus, we trained a classification model on the content to see if it can learn any writing patterns which are not visible to human eye and create benchmark results. Our model, as shown in Figure - 3, is inspired by the works of (Anand et al., 2016) and uses bi-directional LSTM and character embedding to classify articles into questionable (1) and mainstream (0) by learning writing style. We have not performed any preprocessing like removing punctuation, stop words etc. to make sure the model learns every context of the content. 1-D CNN has been used to create character embedding which has been proven to be very efficient in learning orthographic and morphological features of text (Zhang et al., 2015). This layer takes one-hot vector of each character and uses filter size of [196, 196, 300]. We have used 3 layers of 1-D CNN with pool size 2 and kernel stride of 3. Finally, we have performed maxpooling across the sequence to identify features that produces strong sentiments. This layer is then connected to bi-directional LSTM which contains one forward and one backward layer with 128 units each.

This model was trained with top 3 sentences, containing 20-100 words, and retrieved from total 7175 articles. As per the example shown in Table - 2, it can be assumed that top 3 sentences are enough for a normal reader to understand what the article is talking about. Dataset splitting for training and testing was inspired by the findings in the study conducted by (Guess et al., 2019) of Princeton and NYU earlier this year which stated that Americans who shared false stories during 2016 Presidential elections were far more likely to be over 65. Current US age demographic suggests that 15% of the population are over 65 [10]. Therefore, we have decided to have the ratio of questionable to mainstream news articles approximately 1:10, as shown in Table - 3.

### 5.2 Training and Results

We first trained our model by randomly splitting our dataset into training, validation and test set of sizes 4592, 1148 and 1435 respectively. However, since the dataset is unbalanced, stratified k-fold cross-validation mechanism is also implemented with 5 folds. (Kohavi, 1995) states that stratification, which ensures that each fold represents the entire dataset, is generally a better scheme both in terms of bias and variance. The model was trained on an average number of 2, 296, 000 characters. We have evaluated our models on various metrics as shown in Figure - 4. For this problem, we will be majorly focused on ROC and F1 scores. In both the cases, stratified k-fold outperforms random sampling significantly. Training with only random sampling resulted into underrepresentation of data points leading to many false positives. This can also be because writing style of questionable and mainstream articles are very similar. On the other hand, 5-fold cross validation performed significantly well in learning the patterns and avoiding the problems of false positives.

[10] https://www.census.gov/quickfacts/fact/table/US/PST045217

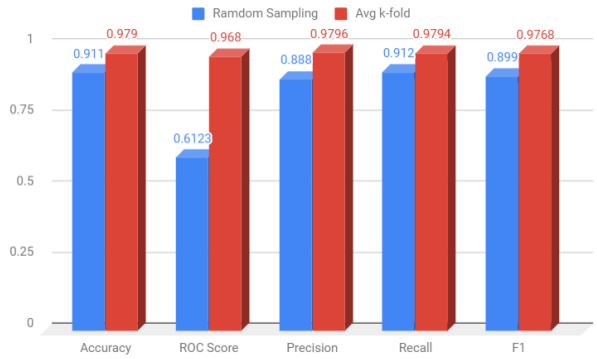


Figure 4: Evaluation results of our model over various metrics. Performance of model using Stratified K-Fold is exceptionally good in terms of ROC score and F1 score.

## 6 Conclusion

In this paper, we have introduced a novel corpus of articles containing false assertions which are written in a very compelling way. We explain how this corpus is different from previously published datasets and explore the characteristics of the corpus. Finally, we use a deep learning classification model to learn the invisible contextual patterns of the content and produce benchmark results on this dataset. Our best model was able to achieve state of the art ROC score of 97%. This is a work in progress and we are planning to extend this corpus by adding more topics and metadata. Future work also involves claim extraction and verification which can be further used to design an unbiased automated detector of intentionally written false stories.

## 7 Acknowledgement

I would like to extend my gratitude to my doctorate advisor, Dr. Rohini K. Srihari, for taking interest into my work and guiding me into the right direction. I would also like to thank my mentor assigned by ACL-SRW, Arkaitz Zubiaga, and anonymous reviewers for their invaluable suggestions and comments on this paper.

## References

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–36.

Ankesh Anand, Tanmoy Chakraborty, and Noseong Park. 2016. We used neural networks to detect clickbaits: You won't believe what happened next! *CoRR*, abs/1612.01340.

Andrew Guess, Jonathan Nagler, and Joshua Tucker. 2019. Less than you think: Prevalence and predictors of fake news dissemination on facebook. *Science Advances*, 5(1).

Ron Kohavi. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'95, pages 1137–1143, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Brendan Nyhan and Jason Reifler. 2010. When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2017. Automatic detection of fake news. *CoRR*, abs/1708.07104.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2017. A stylometric inquiry into hyperpartisan and fake news. *CoRR*, abs/1702.05638.

Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. CSI: A hybrid deep model for fake news. *CoRR*, abs/1703.06959.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and verification. *CoRR*, abs/1803.05355.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. *CoRR*, abs/1705.00648.

Claire Wardle and Hossein Derakhshan. 2017. Information disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe report, DGI (2017)*, 9.

Wenpeng Yin and Dan Roth. 2018. Twowingos: A two-wing optimization strategy for evidential claim verification. *CoRR*, abs/1808.03465.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 649–657. Curran Associates, Inc.