Evons: A Dataset for Fake and Real News Virality Analysis and Prediction

Kriste Krstovski,[1,2] Angela Soomin Ryu,[1] and Bruce Kogut[1]
[1]Columbia Business School, Columbia University
[2]Data Science Institute, Columbia University
{kriste.krstovski,asr2193,bruce.kogut}@columbia.edu

Abstract

We present a novel collection of news articles originating from fake and real news media sources for the analysis and prediction of news virality. Unlike existing fake news datasets which either contain claims or news article headline and body, in this collection each article is supported with a Facebook engagement count which we consider as an indicator of the article virality. In addition we also provide the article description and thumbnail image with which the article was shared on Facebook. These images were automatically annotated with object tags and color attributes. Using cloud based vision analysis tools, thumbnail images were also analyzed for faces and detected faces were annotated with facial attributes. We empirically investigate the use of this collection on an example task of article virality prediction.

1 Introduction

Fake news articles are widely spread across social media platforms such as Facebook and Twitter. This is mainly due to the fact that social media is gradually becoming the main source of news consumption (Shu et al., 2018). Due to the sharing features that these platforms offer, fake news propagate rapidly and their effects resonate and persist across many users (Baly et al., 2018). The wide spread of fake news in social media has lead to the development of automatic fake news detection approaches (Ruchansky et al., 2017; Pérez-Rosas et al., 2018; Nguyen et al., 2019; Zellers et al., 2019), to name a few. Developing fake news detection models require annotated collections of fake and real news articles. Most prior work on the creation and annotation of such collections has focused on this task. Significant number of these collections contain claims fact-checked for veracity (Vlachos and Riedel, 2014; Wang, 2017). A recent survey of such collections is provided in Guo et al. (2022).

On the other hand there exist collections of fake news articles that contain article headline and body text (Horne et al., 2018; Potthast et al., 2018; Zhou et al., 2020). Given that these and other existing fake news collections were developed mainly for fake news detection they can't be used for analysing and predicting fake news virality which is the set of tasks of our focus. Recently, Shu et al. (2018) created FakeNewsNet, a collection of ∼24k news articles labeled as fake or real using the fact-checking websites PolitiFact (PolitiFact, 2017) and Gossip Cop (Gossip Cop, 2021). Articles in this collection are annotated with social engagement information obtained through the Twitter search API. However this collection doesn't include thumbnail images and article descriptions which, along with the headlines, are the only sources of information readers are exposed to on social media platforms regardless of their choices whether to click the link of the shared article or not.

To address this drawback we present Evons – a collection of news articles originating from fake and real news media sources where each article has the thumbnail image and description with which it was shared on Facebook. We use the article engagement count on Facebook as an implicit indicator of the article virality. Given that fake news writers profit from advertising revenue rather than subscription fees, the body text of fake news articles (which are only shown after clicking the link) are known to be repetitive and lacking in informational value (Horne and Adali, 2017). Therefore we believe that these two article components are important for social media sharing. Thumbnail images are annotated with content tags and color attributes while detected faces are annotated with facial attributes. The Evons collection is accessible through `https://github.com/krstovski/evons`. We showcase the use of this collection on the task of article virality prediction which we consider as one example task that could be created wit this dataset.

Proceedings of the 29th International Conference on Computational Linguistics, pages 3589–3596
October 12–17, 2022.

2 Collection Construction

The Evons collection contains 92,969 news articles from fake and real news media sources published in the period between January 2016 and December 2017. We selected this time period to reflect on the 2016 Presidential election which many believed that fake news had a significant impact on. Across both media sources we focused on news articles originating from the same news sections therefore covering similar or the same set of topics. We also don't consider any article author related information. The set of fake news sources was created using information from 3 independent lists of fake news websites that were developed through human curation. It contains only fake news sources that were cross-referenced by at least 2 of the 3 lists. We follow the most widely used definition of fake news as "intentionally and verifiably false, and could mislead readers" by Allcott and Gentzkow (2017) and exclude satire and parody websites.

We used the "Questionable Sources" list from Media Bias Fact Check (MBFC) (Media Bias/Fact Check, 2019) which includes sources with extreme bias, propaganda and conspiracies, and fake news. We filtered the websites to retain only those that are explicitly annotated as "some fake news" or "fake news", indicating that the source deliberately publishes hoaxes and/or disinformation.

Our second list is the "Politifact's Fake News Almanac" (PolitiFact, 2017). This list was created in partnership with Facebook and includes "fake news" websites which were found to contain *deliberately* false or fake stories that have appeared in people's news feeds on Facebook.

The third list is from the "BS Detector" collection. This is a list of "unreliable or otherwise questionable sources" curated by professionals (Risdal, 2017).

After cross-referencing, we obtained 16 fake news sources that appeared in at least 2 of the lists. We then removed sources that were republishing news content from other sources *and* websites that started publishing after the 2016 elections. Our final list contains the following 6 fake news media sources: American Freedom Fighters (AFF), Barracuda Brigade (BB4SP), MadWorldNews (MWN), Puppet String News (PSN), USA Supreme (USAS), and YourNewsWire (YNW). The set of real news sources was obtained from the readily available "All the news 2.0" dataset (Thompson, 2019) which consists of 18 American mainstream sources. We focused on sources that had "high" or "very high" scores in factual reporting *and* "very slight" or "neutral" political biases according to MBFC. There were 5 such sources in this dataset: The Guardian, National Public Radio (NPR), New York Times (NYT), Reuters, and Washington Post (WP). We use articles published in the same time period as our fake media set. In Table 1 we provide the number of articles across the fake and real news media sources.

Media Source	# of Articles
AFF	7,536
BB4SP	2,792
MWN	11,315
PSN	6,576
USAS	3,038
YNW	11,519
Total from fake	42,776
The Guardian	9,811
NPR	11,813
NYT	5,439
Reuters	14,993
WP	8,137
Total from real	50,193
Total	92,969

Table 1: Number of articles in the Evons collection.

We used the webpreview[1] package for extracting thumbnail images. These images come from the thumbnails that are carefully curated by the news producers. They decide what title, description, and thumbnail image would be the most effective in achieving their goal, whether it is to best represent the content or simply attract the most engagement for larger advertising revenue. With this package we also extract article description which is the text that appears as preview when the article is shared.

All articles contain a thumbnail image except for USAS and BB4SP were 0.1% and 11.1% of the articles don't have thumbnails. Thumbnail images are either a picture or a logo of the news media source. Table 2 gives statistics of the number of real and fake articles with and without thumbnail images. Unlike real news articles where a small percentage of them had the media source logo as the thumbnail image, fake news articles always used pictures as thumbnails.

[1] https://pypi.org/project/webpreview

Thumbnail Type	Real	Fake	Total
Picture	48,592	42,464	91,056
Logo	1,601	0	1,601
None	0	312	312

Table 2: Thumbnail statistics.

Engagement Statistics	Real	Fake
Min # of engagements	0	0
Max # of engagements	4.78m	1.08m
Mean # of engagements	6.73	1.58

Table 3: Engagement statistics.

2.1 Engagement Count

A commonly used measure for virality by marketing and communication researchers is how many times a piece of information is shared (Berger and Milkman, 2012; Scholz et al., 2017). Here we use Facebook engagements as a proxy of how much attention the post generated. Facebook engagements is the sum of the number of Facebook shares, likes, and comments. Facebook provides the numbers received by an URL through the Facebook sharing debugger (FSD) (Facebook, 2022). Since FSD works on individual URLs we used the Shared Count API (SharedCount, 2022) to automate the process of fetching these numbers for multiple articles, except for articles from USAS which was blacklisted on Facebook. For this website we used BuzzSumo (BuzzSumo, 2022) which is another third-party measurement dashboard that fetches data from FSD. Both platforms do not provide nor maintain any user related information and have been used in the past across an array of research topics (Xu and Guo, 2018; Xu, 2019; Obiała et al., 2021; Rhodes, 2022). In Table 3 we provide engagement statistics.

2.2 Image Annotation

We performed two types of automatic image annotation. Using Microsoft Azure (Microsoft, 2022) images are analyzed for visual features and color schemes. With the Amazon Rekognition platform (Amazon, 2022) images are analyzed for the presence of faces and detected faces were annotated with facial attributes. Accuracy of both platform on these annotation tasks have been extensively evaluated and confirmed in the past across a variety of image types which include images commonly used as thumbnails (Kyriakou et al., 2019; Liu and Wilkinson, 2020; Malone and Burns, 2021).

Image Tag Statistics	Real	Fake
Min # of tags	0	0
Max # of tags	99	86
Mean # of tags	9.47	9.08

Table 4: Image tag statistics.

2.2.1 Object Detection and Tagging

Images are automatically annotated with content tags such as objects, living beings, scenery, and actions. There were 5,160 distinct tags identified. Articles originating from fake media sources had 3,670 distinct tags with 379 being unique to fake. Real sources contained 4,781 distinct tags with 1,490 unique to real. Table 4 shows image tag statistics. Table 5 shows the top 10 most frequent tags discovered across all media sources, unique to fake, and real news sources.

2.2.2 Color Schemes

Thumbnail images are automatically annotated with three color attributes: dominant foreground and background color, and a set of dominant colors across the whole image. There are 12 colors used: black, blue, brown, gray, green, orange, pink, purple, red, teal, white, and yellow. Dominant background and foreground colors can take on a single value. Thumbnails are also annotated with accent color, which is the most vibrant color in the image, and whether the image is in black and white (bw). In Appendix A we provide summary of the colors present as dominant attribute in thumbnail images.

2.2.3 Facial Analysis

Detected faces are annotated with a bounding box and the following attributes: person's gender, whether the person is smiling, wearing eyeglasses or sunglasses, has a mustache or eyes open, brightness, and sharpness. We also obtain the emotions that appear to be expressed on the face which include: fear, sad, happy, calm, angry, confused, surprised, and disgusted. Table 6 provides face statistics. In Appendix B we show the distribution of dominant face emotions.

3 Example Task

We use the task of predicting article virality as an example task (out of many different tasks) that could be constructed using the Evons collection. The example task is a multi-class classification

All	Unique to Real	Unique to Fake
1. person	1. salad	1. photo caption
2. clothing	2. minimalist	2. television presenter
3. human face	3. raquet sport	3. thong
4. man	4. racketlon	4. shout
5. text	5. piece de resistance	5. g-string
6. outdoor	6. tennis player	6. f-15 eagle
7. suit	7. soft tennis	7. salumi
8. indoor	8. modern	8. salami
9. smile	9. professional boxing	9. ciauscolo
10. tie	10. camera lens	10. ostrich

Table 5: Top 10 most frequent tags across all media sources, unique to real, and fake news sources.

Face Statistics	Real	Fake
% of images with face/s	74.26	77.08
Mean # of faces per image	3.31	2.74

Table 6: Face statistics.

problem which we created by dividing articles from fake and real news media sources into two groups based on their engagement count: real-low, real-high, fake-low, and fake-high. We use the median number of engagements to create almost equal groups of real and fake articles with low and high number of engagements. We empirically investigate how well do various approaches, which we consider as baselines, perform on this task.

3.1 Experimental Setup and Results

The task dataset consists of articles with pictures as thumbnails where the picture contained at least one tag and face. There are 68,793 such articles out of which 36,072 come from real and 32,721 from fake media sources. Articles are represented using two sets of textual features and three sets of image features, one for each of the three image annotation types. For the textual features we use tf-idf values computed over the words of article titles and descriptions. The title feature vector contains 29,745 words and the description feature vector with 43,861 words. Combining both we obtain a vocabulary of 49,792 words. Thumbnail images were represented with 3,526 features: 3,471 object tags, 42 color and 13 facial. Color features include accent color, dominant color attributes, and bw indicator. Facial features include number of faces, person smiling, gender, brightness, sharpness, and facial emotions. Facial features were weighted based on the size of the bounding box area of the detected face. In Appendix C we provide details on the weighing approach used. For features that are indicator variables we use the confidence score as a feature value.

We evaluated 6 different classification models: logistic regression (LR), SVM, multilayer perceptron (MLP), Bidirectional LSTM (Hochreiter and Schmidhuber, 1997) (Bi-LSTM), XLNet (Yang et al., 2019), and RoBERTa (Liu et al., 2019); using a 90/10 split of our dataset. We used the scikit-learn (Pedregosa et al., 2011) implementation of LR and SVM. MLP consists of three fully-connected layers containing 256 and 8 nodes in the first two layers with ReLU. The last layer is a 4 nodes with SoftMax activation. Bi-LSTM consists of a 64 dimensional embedding representation layer, a fully connected layer with ReLU, and an output layer as in MLP. Both NNs were implemented in Keras (Gulli and Pal, 2017). We used the simpletransfomers (Thompson, 2022) implementation of XLNet and RoBERTa with maximum sequence length of 256. Table 7 shows performance comparison results across all models using different feature representations and combinations of them. For ease of interpretability we use accuracy. Thumbnail images were represented using all image generated features. RoBERTa with all feature types performs best. While across most models incorporating image features helps we don't observe substantial accuracy improvement over textual features. We believe that this could be significantly improved with image feature analysis and exploring feature selection approaches.

Feature	Accuracy					
	LR	SVM	MLP	Bi-LSTM	XLNet	RoBERTa
Title (T)	0.632	0.608	0.643	0.632	0.731	0.751
Description (D)	0.674	0.631	0.680	0.687	0.760	0.773
T+D	0.694	0.655	0.718	0.691	0.801	0.807
T+D+Tag	0.701	0.661	0.719	0.712	0.793	0.808
T+D+Color	0.701	0.658	0.716	0.688	0.781	0.801
T+D+Facial	0.697	0.655	0.716	0.688	0.794	0.802
All	0.703	0.666	0.714	0.683	0.791	0.810

Table 7: Accuracy results across various baseline models on the example task of article virality prediction.

4 Conclusion

We presented Evons - a collection of news articles originating from fake and real media sources where articles are annotated with a Facebook engagement count, thumbnail image and article description. Thumbnails are automatically annotated with object tags, color and facial attributes. We demoed the collection use on an article virality prediction task and established baselines using 6 models. In the future we plan to use Evons to explore various approaches for selection of image features and combination with text that would further help improve accuracy on this task.

5 Ethics

Creating the Evons collection involved collecting news articles from various online media sources, extracting thumbnail images using the webpreview package, and obtaining Facebook engagement counts through the SharedCount API and the BuzzSumo platforms. Throughout the creation process we made sure that no author metadata or user identifying information was collected. Therefore our collection does not contain any information that names or uniquely identifies individual people. Both Facebook engagement counts platforms do not provide any user related information. While news articles across various online media sources do provide article author information in our collection process we ignored this information.

We don't foresee any potential risks that may arise from the creation of our collection especially in terms of identifying potential stakeholders that may benefit from this collection while harming others. To the best of our knowledge all of our collected data is in the public domain and is not copyrighted.

For our thumbnail image annotations we relied on two image annotation platforms: Microsoft Azure and Amazon Rekognition. One limitation of our work may arise from the fact that we don't know whether the models that are part of these platforms contain any type of bias and if so to which extent bias is present.

References

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

Amazon. 2022. Detecting and analyzing faces. https://docs.aws.amazon.com/rekognition/latest/dg/faces [Accessed: April, 2022].

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. In *EMNLP*, pages 3528–3539.

Jonah Berger and Katherine L. Milkman. 2012. What makes online content viral? *Journal of Marketing Research*, 49(2):192–205.

BuzzSumo. 2022. https://buzzsumo.com [Accessed: April, 2022].

Facebook. 2022. https://developers.facebook.com/tools/debug/ [Accessed: April, 2022].

Gossip Cop. 2021. https://www.gossipcop.com/about.html [Accessed: October, 2021].

Antonio Gulli and Sujit Pal. 2017. *Deep learning with Keras*. Packt Publishing Ltd.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Benjamin D Horne and Sibel Adali. 2017. This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *AAAI Conference on Web and Social Media*.

Benjamin D Horne, Sara Khedr, and Sibel Adali. 2018. Sampling the news producers: A large news and feature data set for the study of the complex media landscape. In *AAAI Conference on Web and Social Media*.

Kyriakos Kyriakou, Pınar Barlas, Styliani Kleanthous, and Jahna Otterbacher. 2019. Fairness in proprietary image tagging algorithms: A cross-platform audit on people images. In *AAAI Conference on Web and Social Media*, volume 13, pages 313–322.

Ching Yiu Jessica Liu and Caroline Wilkinson. 2020. Image conditions for machine-based face recognition of juvenile faces. *Science & Justice*, 60(1):43–52.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ashling Malone and John Burns. 2021. Evaluating the accuracy of public cloud vendor face detection api's. *Journal of Image and Graphics*, 9(1).

Media Bias/Fact Check. 2019. Media bias/fact check questionable sources. `https://mediabiasfactcheck.com/fake-news/` [Accessed: December, 2019].

Microsoft. 2022. What is computer vision? `https://docs.microsoft.com/en-us/azure/cognitive-services/computer-vision/home` [Accessed: April, 2022].

Duc Minh Nguyen, Tien Huu Do, Robert Calderbank, and Nikos Deligiannis. 2019. Fake news detection using deep Markov random fields. In *NAACL*, pages 1391–1400.

Justyna Obiała, Karolina Obiała, Małgorzata Mańczak, Jakub Owoc, and Robert Olszewski. 2021. Covid-19 misinformation: accuracy of articles about coronavirus prevention mostly shared on social media. *Health policy and technology*, 10(1):182–186.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *JMLR*, 12:2825–2830.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *COLING*, pages 3391–3401.

PolitiFact. 2017. Politifact's guide to fake news websites and what they peddle. `https://www.politifact.com/article/2017/apr/20/politifacts-guide-fake-news-websites-and-what-they` [Accessed: October, 2021].

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylometric inquiry into hyperpartisan and fake news. In *ACL*, pages 231–240.

Samuel C Rhodes. 2022. Filter bubbles, echo chambers, and fake news: how social media conditions individuals to be less critical of political misinformation. *Political Communication*, 39(1):1–22.

Megan Risdal. 2017. Getting Real about Fake News. `https://www.kaggle.com/mrisdal/fake-news` [Accessed: April, 2022].

Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. Csi: A hybrid deep model for fake news detection. In *CIKM*, pages 797–806.

Christin Scholz, Elisa C. Baek, Matthew Brook O'Donnell, Hyun Suk Kim, Joseph N. Cappella, and Emily B. Falk. 2017. A neural model of valuation and information virality. *PNAS*, 114(11):2881–2886.

SharedCount. 2022. `https://www.sharedcount.com` [Accessed: April, 2022].

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*.

Andrew Thompson. 2019. "all the news 2.0" dataset. `https://components.one/datasets/all-the-news-articles-dataset` [Accessed: December, 2019].

Andrew Thompson. 2022. `https://simpletransformers.ai` [Accessed: April, 2022].

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *ACL*, pages 18–22.

William Yang Wang. 2017. ”liar, liar pants on fire”: A new benchmark dataset for fake news detection. In *ACL*, pages 422–426.

Zhan Xu. 2019. Personal stories matter: topic evolution and popularity among pro-and anti-vaccine online articles. *Journal of computational social science*, 2(2):207–220.

Zhan Xu and Hao Guo. 2018. Using text mining to compare online pro-and anti-vaccine headlines: word usage, sentiments, and online popularity. *Communication Studies*, 69(1):103–122.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*, pages 5753–5763.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *NeurIPS*, pages 9054–9065.

Xinyi Zhou, Apurva Mulay, Emilio Ferrara, and Reza Zafarani. 2020. Recovery: A multimodal repository for covid-19 news credibility research. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 3205–3212.

A Dominant Colors

Shown in Figure 1 are bar plots of the percentage of colors present as dominant attribute in thumbnail images.

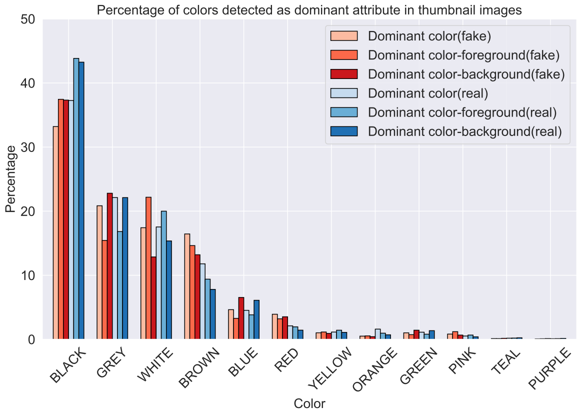

Figure 1: Percentage of color present as dominant attribute in thumbnail images.

B Dominant Emotions

Shown in Figure 2 are bar plots of the percentage of emotion detected as dominant on faces found in thumbnail images.

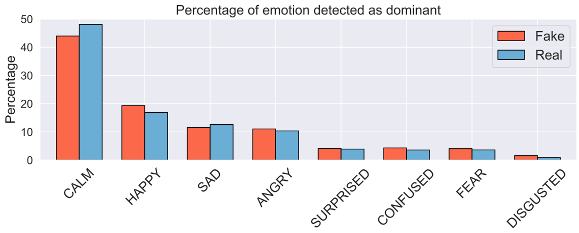

Figure 2: % of emotion detected as dominant in faces.

C Facial Features

Facial features across thumbnail images where weighted based on the bounding box area of the detected face. The bounding box area is the product of the bounding box width and height. Given a bounding box area B_{ij} of the jth face in image i and a set of k features F_{jk} detected on that face, the weighted facial features for image i, W_{ik} are computed as:

$$W_{ik} = \sum_{j=1}^{J} B_{i,j} F_{j,k} \tag{1}$$