r/Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection

Kai Nakamura
Laguna Blanca School
Santa Barbara, CA 93110
kai.nakamura42@gmail.com

Sharon Levy, William Yang Wang
Department of Computer Science
University of California, Santa Barbara
Santa Barbara, CA 93106 USA
{sharonlevy,william}@cs.ucsb.edu

Abstract

Fake news has altered society in negative ways as evidenced in politics and culture. It has adversely affected both online social network systems as well as offline communities and conversations. Using automatic fake news detection algorithms is an efficient way to combat the rampant dissemination of fake news. However, using an effective dataset has been a problem for fake news research and detection model development. In this paper, we present Fakeddit, a novel dataset consisting of about 800,000 samples from multiple categories of fake news. Each sample is labeled according to 2-way, 3-way, and 5-way classification categories. Prior fake news datasets do not provide multimodal text and image data, metadata, comment data, and fine-grained fake news categorization at this scale and breadth. We construct hybrid text+image models and perform extensive experiments for multiple variations of classification.

1 Introduction

Within our progressively digitized society, the spread of fake news and misinformation has enlarged, leading to many problems such as an increasingly politically divisive climate. The dissemination and consequences of fake news are exacerbating partly due to the rise of popular social media applications with inadequate fact-checking or third-party filtering, enabling any individual to broadcast fake news easily and at a large scale (Allcott and Gentzkow, 2017). Though steps have been taken to detect and eliminate fake news, it still poses a dire threat to society (Dreyfuss and Lapowsky). As such, research in the area of fake news detection is essential.

To build any machine learning model, one must obtain good training data for the specified task. In the realm of fake news detection, there are several existing published datasets. However, they have several limitations: limited size, modality, and/or granularity. Though fake news may immediately be thought of as taking the form of text, it can appear in other mediums such as images. As such, it is important that standard fake news detection systems detect all types of fake news and not just text data. Our dataset will expand fake news research into the multimodal space and allow researchers to develop stronger fake news detection systems.

Our contributions to the study of fake news detection are:

- We create a large-scale multimodal fake news dataset consisting of around 800,000 samples containing text, image, metadata, and comments data from a highly diverse set of resources.
- Each data sample consists of multiple labels, allowing users to utilize the dataset for 2-way, 3-way, and 5-way classification. This enables both high-level and fine-grained fake news classification.
- We evaluate our dataset through text, image, and text+image modes with a neural network architecture that integrates both the image and text data. We run experiments for several types of models, providing a comprehensive overview of classification results.

2 Related Work

A variety of datasets for fake news detection have been published in recent years. These are listed in Table 1, along with their specific characteristics. When comparing these datasets, a few trends can be seen. Most of the datasets are small in size, which can be ineffective for current machine learning models that require large quantities of training data. Only four contain over half a million samples, with CREDBANK and FakeNewsCorpus[1] being the largest with millions of sam-

[1]https://github.com/several27/FakeNewsCorpus

Dataset	Size	Number of Classes	Modality	Source	Data Category
LIAR	12,836	6	text	Politifact	political
FEVER	185,445	3	text	Wikipedia	variety
BUZZFEEDNEWS	2,282	4	text,image	Facebook	political
BUZZFACE	2,263	4	text,image	Facebook	political
some-like-it-hoax	15,500	2	text	Facebook	scientific/conspiracy
PHEME	330	2	text,image	Twitter	variety
CREDBANK	60,000,000	5	text	Twitter	variety
Breaking!	700	2,3	text	BS Detector	political
NELA-GT-2018	713,000	8 IA	text	194 news outlets	variety
FAKENEWSNET	602,659	2	text,image	Twitter	political/celebrity
FakeNewsCorpus	9,400,000	10	text	Opensources.co	variety
FA-KES	804	2	text	15 news outlets	Syrian war
image-verification-corpus	15,629	2	text,image	Twitter	variety
Image Manipulation	48	2	image	self-taken	variety
Fakeddit	825,100	2,3,5	image,text	Reddit	variety

Table 1: Comparison of various fake news detection datasets. IA: Individual assessments.

ples (Mitra and Gilbert, 2015). In addition, many of the datasets separate their data into a small number of classes, such as fake vs. true. However, fake news can be categorized into many different types (Wardle). Datasets such as NELA-GT-2018, LIAR, and FakeNewsCorpus provide more fine-grained labels (Nrregaard et al., 2019; Wang, 2017). While some datasets include data from a variety of categories (Zubiaga et al., 2016; Thorne et al., 2018), many contain data from specific areas, such as politics and celebrity gossip (Tacchini et al., 2017; Pathak and Srihari, 2019; Shu et al., 2018; Abu Salem et al., 2019). These data samples may contain limited styles of writing due to this categorization. Finally, most of the existing fake news datasets collect only text data, which is not the only mode that fake news can appear in. Datasets such as image-verification-corpus, Image Manipulation, BUZZFEEDNEWS[2], and BUZZFACE can be utilized for fake image detection, but contain small sample sizes(Christlein et al., 2012; Boididou et al., 2018; Santia and Williams, 2018). It can be seen from the table that compared to other existing datasets, Fakeddit contains a large quantity of data, while also annotating for three different types of classification labels (2-way, 3-way, and 5-way) and comparing both text and image data.

3 Fakeddit

Many fake news datasets are crowdsourced or handpicked from a select few sources that are narrow in size, modality, and/or diversity. In order to expand and evolve fake news research, researchers need to have access to a dataset that exceed these current dataset limitations. Thus, we propose Fakeddit[3], a novel dataset consisting of a large quantity of text+image samples coming from large diverse sources.

We sourced our dataset from Reddit[4], a social news and discussion website where users can post submissions on various subreddits. Each subreddit has its own theme like 'nottheonion'[5], where people post seemingly false stories that are surprisingly true. Active Reddit users are able to upvote, downvote, and comment on the submission.

Submissions were collected with the pushshift.io API[6]. Each subreddit has moderators that ensure submissions pertain to the subreddit theme and remove posts that violate any rules, indirectly helping us obtain reliable data. To further ensure that our data is credible, we filtered out any submissions that had a score of less than 1. Fakeddit consists of 825,100 total submissions from 21 different subreddits. We gathered the submission title and image, comments made by users who engaged with the submission, as well as other submission metadata including the score, the username of the author, subreddit source, sourced domain, number of comments, and up-vote to down-vote ratio. 63% of the samples contains both text and images, while the rest contain only text. For our experiments, we utilize these multimodal samples. The samples span over many years and are posted on highly active and

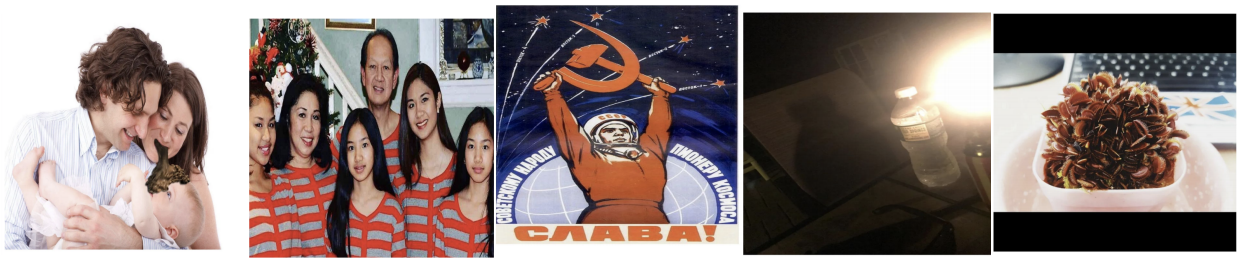
[2]https://github.com/BuzzFeedNews/2016-10-facebook-fact-check

[3]https://github.com/entitize/fakeddit

[4]https://www.reddit.com/

[5]https://www.reddit.com/r/nottheonion

[6]https://pushshift.io/



(a) Satire/Parody (b) True (c) Misleading Content (d) Imposter Content (e) False Connection

Figure 1: Dataset examples with 5-way classification labels.

popular pages by tens of thousands of diverse individual users from across the world. Because of the variety of the chosen subreddits, our data also varies in its content, ranging from political news stories to simple everyday posts by Reddit users.

We provide three labels for each sample, allowing us to train for 2-way, 3-way, and 5-way classification. Having this hierarchy of labels will enable researchers to train for fake news detection at a high level or a more fine-grained one. The 2-way classification determines whether a sample is fake or true. The 3-way classification determines whether a sample is completely true, the sample is fake news with true text (text that is true in the real world), or the sample is fake news with false text. Our final 5-way classification was created to categorize different types of fake news rather than just doing a simple binary or trinary classification. This can help in pinpointing the degree and variation of fake news for applications that require this type of fine-grained detection. The first label is true and the other four are defined within the seven types of fake news (Wardle). We provide examples from each class for 5-way classification in Figure 3. The 5-way classification labels are explained below:

True: True content is accurate in accordance with fact. Eight of the subreddits fall into this category, such as usnews[7] and mildlyinteresting[8]. The former consists of posts from various news sites. The latter encompasses real photos with accurate captions. The other subreddits include photoshopbattles[9], nottheonion, neutralnews[10], pic[11], usanews[12], and upliftingnews[13].

Satire/Parody: This category consists of content that spins true contemporary content with a satirical tone or information that makes it false. One of the four subreddits that make up this label is theonion[14], with headlines such as "Man Lowers Carbon Footprint By Bringing Reusable Bags Every Time He Buys Gas". Other satirical subreddits are fakealbumcovers[15], satire[16], and waterfordwhispersnews[17].

Misleading Content: This category consists of information that is intentionally manipulated to fool the audience. Our dataset contains three subreddits in this category: propagandaposters[18], fakefacts[19], and savedyouaclick[20].

Imposter Content: This category contains the subredditsimulator[21] subreddit, which contains bot-generated content and is trained on a large number of other subreddits. It also includes subsimulatorgpt2[22].

False Connection: Submission images in this category do not accurately support their text descriptions. We have four subreddits with this label, containing posts of images with captions

[7] https://www.reddit.com/r/usnews
[8] https://www.reddit.com/r/mildlyinteresting
[9] https://www.reddit.com/r/photoshopbattles
[10] https://www.reddit.com/r/neutralnews
[11] https://www.reddit.com/r/pic
[12] https://www.reddit.com/r/usanews
[13] https://www.reddit.com/r/upliftingnews
[14] https://www.reddit.com/r/theonion
[15] https://www.reddit.com/r/fakealbumcovers
[16] https://www.reddit.com/r/satire
[17] https://www.reddit.com/r/waterfordwhispersnews
[18] https://www.reddit.com/r/propagandaposters
[19] https://www.reddit.com/r/fakefacts
[20] https://www.reddit.com/r/savedyouaclick
[21] https://www.reddit.com/r/subredditsimulator
[22] https://www.reddit.com/r/subsimulatorgpt2

		2-way			3-way		5-way	
Type	Text	Image	Validation	Test	Validation	Test	Validation	Test
Text	BERT	–	0.769	0.773	0.761	0.767	0.741	0.739
	InferSent	–	**0.783**	**0.786**	**0.776**	**0.779**	**0.746**	**0.746**
Image	–	VGG16	0.695	0.698	0.678	0.677	0.638	0.636
	–	EfficientNet	0.560	0.561	0.560	0.561	0.547	0.545
	–	ResNet50	**0.721**	**0.722**	**0.712**	**0.711**	**0.675**	**0.673**
Text+Image	InferSent	VGG16	0.841	0.839	0.829	0.831	0.808	0.806
	InferSent	EfficientNet	0.787	0.788	0.780	0.784	0.749	0.746
	InferSent	ResNet50	0.857	0.854	0.850	0.849	0.819	0.818
	BERT	VGG16	0.846	0.846	0.837	0.837	0.810	0.809
	BERT	EfficientNet	0.787	0.788	0.780	0.783	0.746	0.746
	BERT	ResNet50	**0.863**	**0.863**	**0.859**	**0.859**	**0.832**	**0.830**

Table 2: Results on fake news detection for 2, 3, and 5-way classification with combination method of maximum.

	2-way		3-way		5-way	
Combination Methods	Validation	Test	Validation	Test	Validation	Test
Add	0.810	0.814	0.797	0.799	0.786	0.783
Concatenate	0.812	0.814	0.807	0.809	0.787	0.783
Maximum	**0.863**	**0.863**	**0.859**	**0.859**	**0.832**	**0.830**
Average	0.816	0.816	0.811	0.813	0.801	0.795

Table 3: Results on different multi-modal combinations for BERT + ResNet50

that do not relate to the true meaning of the image. These include misleadingthumbnails[23], confusing_perspective[24], pareidolia[25], and fakehistoryporn[26].

4 Experiments

4.1 Fake News Detection

Multiple methods were employed for text and image feature extraction. We used InferSent and BERT to generate text embeddings for the title of the Reddit submissions (Conneau et al., 2017; Devlin et al., 2018). VGG16, EfficientNet, and ResNet50 were utilized to extract the features of the Reddit submission thumbnails (Simonyan and Zisserman, 2015; Tan and Le, 2019; He et al., 2015).

We used the InferSent model because it performs very well as a universal sentence embeddings generator. For this model, we loaded a vocabulary of 1 million of the most common words in English and used fastText as opposed to ELMO embeddings because fastText can perform relatively well for rare words and words that do not appear in the vocabulary (Joulin et al., 2016; Peters et al., 2018). We obtained encoded sentence features of length 4096 for each submission title using InferSent.

The BERT model achieves state-of-the-art results on many classification tasks, including Q&A and named entity recognition. To obtain fixed-length BERT embedding vectors, we used the bert-as-service tool, which maps variable-length text/sentences into a 768 element array for each Reddit submission title (Xiao, 2018). For our experiments, we utilized the pretrained BERT-Large, Uncased model.

We utilized VGG16, ResNet50, and EfficientNet models for encoding images. VGG16 and ResNet50 are widely used by many researchers, while EfficientNet is a relatively newer model. For EfficientNet, we used the smallest variation: B0. For all three image models, we preloaded weights of models trained on ImageNet and included the top layer and used its penultimate layer for feature extraction.

For our experiments, we excluded submissions that did not have an image associated with them and solely used submission image and title data. We performed 2-way, 3-way, and 5-way classification for each of the three types of inputs: image only, text only, and multimodal (text and image).

Before training, we performed preprocessing on the images and text. We constrained sizes of the images to 224x224. From the text, we removed all punctuation, numbers, and revealing words such as PsBattle that automatically reveal the subreddit source. For the savedyouaclick subreddit, we

[23]https://www.reddit.com/r/misleadingthumbnails
[24]https://www.reddit.com/r/confusing_perspective
[25]https://www.reddit.com/r/pareidolia
[26]https://www.reddit.com/r/fakehistoryporn

removed text following the "|" character and classified it as misleading content.

When combining the features in multimodal classification, we first condensed the features into 256-element vectors through a trainable dense layer and then merged them through four different methods: add, concatenate, maximum, average. These features were then passed through a fully connected softmax predictor.

4.2 Results

The results are shown in Tables 2 and 3. We found that the multimodal features performed the best, followed by text-only, and image-only in all instances. Thus, having both image and text improves fake news detection. For image and multimodal classification, ResNet50 performed the best followed by VGG16 and EfficientNet. In addition, BERT generally achieved better results than InferSent for multimodal classification. However, for text-only classification InferSent outperformed BERT. The maximum method to merge image and text features yielded the highest accuracy, followed by average, concatenate, and add. Overall, the multimodal model that combined BERT text features and ResNet50 image features through the maximum method performed most optimally.

5 Conclusion

In this paper, we presented a novel dataset for fake news research, Fakeddit. Compared to previous datasets, Fakeddit provides a large quantity of text+image samples with multiple labels for various levels of fine-grained classification. We created detection models that incorporate both modalities of data and conducted experiments, showing that there is still room for improvement in fake news detection. Although we do not utilize submission metadata and comments made by users on the submissions, we anticipate that these features will be useful for further research. We hope that our dataset can be used to advance efforts to combat the ever growing rampant spread of misinformation.

Acknowledgments

We would like to acknowledge Facebook for the Online Safety Benchmark Award. The authors are solely responsible for the contents of the paper, and the opinions expressed in this publication do not reflect those of the funding agencies.

References

Fatima K. Abu Salem, Roaa Al Feel, Shady Elbassuoni, Mohamad Jaber, and May Farah. 2019. Fa-kes: A fake news dataset around the syrian war. *Proceedings of the International AAAI Conference on Web and Social Media*, 13(01):573–582.

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

Christina Boididou, Symeon Papadopoulos, Markos Zampoglou, Lazaros Apostolidis, Olga Papadopoulou, and Yiannis Kompatsiaris. 2018. Detection and visualization of misleading content on twitter. *International Journal of Multimedia Information Retrieval*, 7(1):71–86.

Vincent Christlein, Christian Riess, Johannes Jordan, Corinna Riess, and Elli Angelopoulou. 2012. An evaluation of popular copy-move forgery detection approaches. *IEEE Transactions on information forensics and security*, 7(6):1841–1854.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Emily Dreyfuss and Issie Lapowsky. Facebook is changing news feed (again) to stop fake news. *Wired*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Tanushree Mitra and Eric Gilbert. 2015. Credbank: A large-scale social media corpus with associated credibility annotations. In *Ninth International AAAI Conference on Web and Social Media*.

Jeppe Nrregaard, Benjamin D. Horne, and Sibel Adal. 2019. Nela-gt-2018: A large multi-labelled news dataset for the study of misinformation in news articles. *Proceedings of the International AAAI Conference on Web and Social Media*, 13(01):630–638.

Archita Pathak and Rohini Srihari. 2019. BREAKING! presenting fake news corpus for automated fact checking. In *Proceedings of the 57th Annual*

Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 357–362, Florence, Italy. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Giovanni C Santia and Jake Ryland Williams. 2018. Buzzface: A news veracity dataset with facebook user commentary and egos. In *Twelfth International AAAI Conference on Web and Social Media*.

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*.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

Eugenio Tacchini, Gabriele Ballarin, Marco L. Della Vedova, Stefano Moret, and Luca de Alfaro. 2017. Some like it hoax: Automated fake news detection in social networks. *CoRR*, abs/1704.07506.

Mingxing Tan and Quoc V. Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Claire Wardle. Fake news. its complicated. *First Draft*.

Han Xiao. 2018. bert-as-service. `https://github.com/hanxiao/bert-as-service`.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 11(3):e0150989.