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

Kai Nakamura*¶, Sharon Levy*§, William Yang Wang§

¶Laguna Blanca School
§University of California, Santa Barbara
kai.nakamura42@gmail.com, {sharonlevy, william}@cs.ucsb.edu

Abstract

Fake news has altered society in negative ways in politics and culture. It has adversely affected both online social network systems as well as offline communities and conversations. Using automatic machine learning classification models is an efficient way to combat the widespread dissemination of fake news. However, a lack of effective, comprehensive datasets has been a problem for fake news research and detection model development. Prior fake news datasets do not provide multimodal text and image data, metadata, comment data, and fine-grained fake news categorization at the scale and breadth of our dataset. We present Fakeddit, a novel multimodal dataset consisting of over 1 million samples from multiple categories of fake news. After being processed through several stages of review, the samples are labeled according to 2-way, 3-way, and 6-way classification categories through distant supervision. We construct hybrid text+image models and perform extensive experiments for multiple variations of classification, demonstrating the importance of the novel aspect of multimodality and fine-grained classification unique to Fakeddit.

Keywords: fake news, machine learning, multimodal

1. Introduction

Within our progressively digitized society, the spread of fake news and disinformation has enlarged in journalism, news reporting, social media, and other forms of online information consumption. False information from these sources, in turn, has caused many problems such as spurring irrational fears during medical outbreaks like Ebola[1]. The dissemination and consequences of fake news are exacerbating due to the rise of popular social media applications and other online sources 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, 2019). According to a Pew Research Center report[2], 50% of Americans view fake news as a critical problem, placing it above violent crime. In addition, the report found that 68% of Americans view fake news as having a significant impact on their confidence of the government and 54% viewed it as having a large impact in their trust in one another. As such, research in the area of fake news detection is of high importance for society.

To build a fake news detection model, one must obtain sizable and diverse training data. Within this area of research, there are several existing published datasets. However, they have many constraints: limited size, modality, and granularity. Most conventional fake news research and datasets such as LIAR (Wang, 2017) and Some-Like-It-Hoax (Tacchini et al., 2017) solely focus on text data. However, online information today is also consumed through multime-

dia sources including images, which often supplement the text. In addition, many datasets are small in size and variation. For example, Abu Salem et al. (2019) aims to increase the diversity of fake news by covering news that goes beyond the scope of conventional American political news. However, it suffers the problem of only consisting of less than 1000 samples, limiting its extent to which it can contribute to fake news research. Moreover, many conventional datasets label their data binarily (true and false). However, fake news can be categorized into many different types (Wardle, 2017). These problems significantly affect the quality of fake news research and detection.

We overcome these limitations posed by conventional datasets through the dataset we propose: Fakeddit[3], a novel multimodal fake news detection dataset consisting of over 1 million samples with 2-way, 3-way, and 6-way classification labels, along with comment data and metadata. We sourced our data from multiple subreddits from Reddit[4]. Our dataset will expand fake news research into the multimodal space and allow researchers to develop stronger, more generalized, fine-grained fake news detection systems. We provide examples from our dataset in Figure 1. Our contributions to the study of fake news detection are:

- We create a large-scale multimodal fake news dataset consisting of over 1 million 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 6-way classification. This enables both high-level and fine-grained fake news classification. Samples are also

---

* Equal Contribution.

[1] https://www.pbs.org/newshour/science/real-consequences-fake-news-stories-brain-cant-ignore

[2] https://www.journalism.org/2019/06/05/many-americans-say-made-up-news-is-a-critical-problem-that-needs-to-be-fixed/

[3] `https://github.com/entitize/fakeddit`
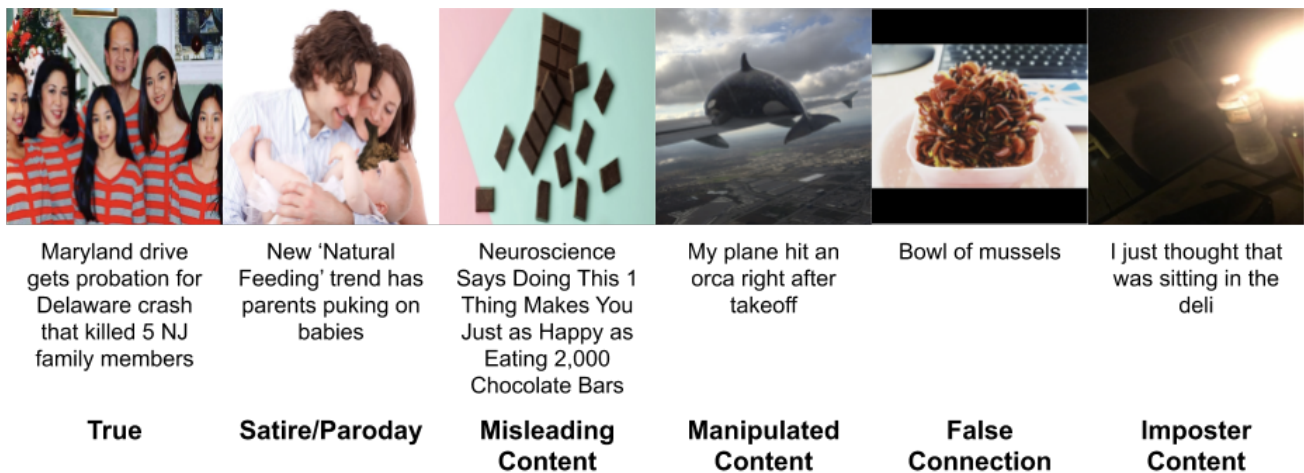[4] http://reddit.com/

Figure 1: Dataset examples with 6-way classification labels.

thoroughly refined through multiple steps of quality assurance.

- We evaluate our dataset through text, image, and text+image modes with neural network architectures that integrate both the image and text data. We run experiments for several types of baseline models, providing a comprehensive overview of classification results and demonstrating the significance of multimodality present in Fakeddit.

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.

2.1. Text Datasets

When comparing fake news 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 datasets contain over half a million samples, with CREDBANK (Mitra and Gilbert, 2015) and FakeNewsCorpus[5] being the largest, both containing millions of samples. In addition, many of the datasets separate their data into a small number of classes, such as fake vs. true. Datasets such as NELA-GT-2018 (Nørregaard et al., 2019), LIAR (Wang, 2017), and FakeNewsCorpus provide more fine-grained labels. While some datasets include data from a variety of categories (Zubiaga et al., 2016), 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; Santia and Williams, 2018)[6]. These data samples may contain limited scopes of context and styles of writing due to their limited number of categories.

---

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

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

2.2. Image Datasets

Most of the existing fake news datasets collect only text data. However, fake news can also come in the form of images. Existing fake image datasets are limited in size and diversity, making dataset research in this area important. Image features supply models with more data that can help immensely to identify fake images and news that have image data. We analyze three traditional fake image datasets that have been published. The Image Manipulation dataset (Christlein et al., 2012) contains self-taken manipulated images for image manipulation detection. The PS-Battles dataset (Heller et al., 2018) is an image dataset containing manipulated image derivatives from one subreddit. We expand upon the size and scope of the data provided from the same subreddit in the PS-Battles dataset by expanding the size and time range as well as including text data and other metadata. This expanded data makes up only two of 22 sources of data present in our research. The image-verification-corpus (Boididou et al., 2018), like ours, contains both text and image data. While it does contain a larger number of samples than other conventional datasets, it still pales in comparison to Fakeddit.

2.3. Fact-Checking

Due to the unique aspect of multimodality, Fakeddit can also be applied to the realm of implicit fact-checking. Other existing datasets utilized for fact-checking include FEVER (Thorne et al., 2018) and Fauxtography (Zlatkova et al., 2019). The former consists of altered claims utilized for textual verification. The latter utilizes both text and image data in order to fact-check claims about images. Using both text and image data, researchers can use Fakeddit for verifying truth and proof: utilizing image data as evidence for text truthfulness or using the text data as evidence for image truthfulness.

Compared to other existing datasets, Fakeddit provides a larger breadth of novel features that can be applied in a number of applications: fake news text, image, text+image classification as well as implicit fact-checking. Other data provided, such as comments data, enable more applications.

Dataset	Size (# of samples)	# 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	Facebook	political
BUZZFACE	2,263	4	text	Facebook	political
some-like-it-hoax	15,500	2	text	Facebook	scientific/conspiracy
PHEME	330	2	text	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	Twitter	political/celebrity
FakeNewsCorpus	9,400,000	10	text	Opensources.co	variety
FA-KES	804	2	text	15 news outlets	Syrian war
Image Manipulation	48	2	image	self-taken	variety
Fauxtography	1,233	2	text, image	Snopes, Reuters	variety
image-verification-corpus	17,806	2	text, image	Twitter	variety
The PS-Battles Dataset	102,028	2	image	Reddit	manipulated content
Fakeddit (ours)	**1,063,106**	**2,3,6**	**text, image**	**Reddit**	**variety**

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

3. Fakeddit

3.1. Data Collection

We sourced our dataset from Reddit, a social news and discussion website where users can post submissions on various subreddits. Reddit is one of the top 20 websites in the world by traffic[7]. Each subreddit has its own theme. For example, 'nottheonion' is a subreddit where people post seemingly false stories that are surprisingly true. Active Reddit users are able to upvote, downvote, and submit comments on the submissions.

Fakeddit consists of over 1 million submissions from 22 different subreddits. The specific subreddits can be found in the Appendix. As depicted in Table 2, the samples span over almost a decade and are posted on highly active and popular pages by over 300,000 unique individual users, allowing us to capture a wide variety of perspectives. Having a decade's worth of recent data allows machine learning models to stay attuned to contemporary cultural-linguistic patterns and current events. Our data also varies in its content, because of the array of the chosen subreddits, ranging from political news stories to simple everyday posts by Reddit users.

Submissions were collected with the pushshift.io API[8] with the earliest submission being from March 19, 2008, and the most recent submission being from October 24, 2019. 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. From the photoshopbattles subreddit, we treated both submission and comment data as submission data. In the photoshopbattles subreddit, users post submissions that contain true images. Other users then manipulate these submission images and post these doctored images as comments on the submission's page. These comments also contain text data that

Dataset Statistics	
Total samples	1,063,106
Fake samples	628,501
True samples	527,049
Multimodal samples	682,996
Subreddits	22
Unique users	358,504
Unique domains	24,203
Timespan	3/19/2008 - 10/24/2019
Mean words per submission	8.27
Mean comments per submission	17.94
Vocabulary size	175,566
Training set size	878,218
Validation set size	92,444
Released test set size	92,444
Unreleased set size	92,444

Table 2: Fakeddit dataset statistics

relate or describe the image. We harvest these comments from the photoshopbattles subreddit and treat them as submission data to incorporate in our submission dataset, significantly contributing to the total number of multimodal samples. Approximately 64% of the samples in our dataset contain both text and images. These multimodal samples are used for our baseline experiments and error analysis.

3.2. Quality Assurance

Because our dataset contains over one million samples, it is crucial to make sure that it contains reliable data. To do so, we have several levels of data processing. The first is provided through the subreddit pages. Each subreddit has moderators that ensure submissions pertain to the subreddit theme. The job of these moderators is to remove posts that violate any rules. As a result, the data goes through its first round of refinement. The next stage occurs when we start collecting the data. In this phase, we utilize Reddit's upvote/downvote score feature. This feature is intended to not only signify another user's approval for the post but also

---

[7]https://www.alexa.com/topsites
[8]https://pushshift.io/

indicates that a post does not contribute to the subreddit's theme or is off-topic if it has a low score[9]. As such, we filtered out any submissions that had a score of less than 1 to further ensure that our data is credible. We assume that invalid or irrelevant posts within a subreddit would be either removed or down-voted to a score of less than 1. The high popularity of the Reddit website makes this step particularly effective as thousands of individual users can give their opinion of the quality of various submissions.

Our final degree of quality assurance is done manually and occurs after the previous two stages. We randomly sampled 10 posts from each subreddit in order to determine whether the submissions really do pertain to each subreddit's theme. If any of the 10 samples varied from this, we decided to remove the subreddit from our list. As a result, we ended up with 22 subreddits to keep our processed data after this filtering. When labeling our dataset, we labeled each sample according to its subreddit's theme. These labels were determined during the last processing phase, as we were able to look through many samples for each subreddit. Each subreddit is labeled with one 2-way, 3-way, and 6-way label. Lastly, we cleaned the submission title text: we removed all punctuation, numbers, and revealing words such as "PsBattle" and "colorized" that automatically reveal the subreddit source. For the savedyouaclick subreddit, we removed text following the "|" character and classified it as misleading content. We also converted all the text to lowercase.

As mentioned above, we do not manually label each sample and instead label our samples based on their respective subreddit's theme. By doing this, we employ distant supervision, a commonly used technique, to create our final labels. While this may create some noise within the dataset, we aim to remove this from our pseudo-labeled data. By going through these stages of quality assurance, we can determine that our final dataset is credible and each subreddit's label will accurately identify the posts that it contains. We test this by randomly sampling 150 text-image pairs from our dataset and having two of our researchers individually manually label them for 6-way classification. It is difficult to narrow down each sample to exactly one subcategory, especially for those not working in the journalism industry. We achieve a Cohen's Kappa coefficient (Cohen, 1960) of 0.54, showing moderate agreement and that some samples may represent more than one label. While we only provide each sample with one 6-way label, future work can help identify multiple labels for each text-image pair.

3.3. Labeling

We provide three labels for each sample, allowing us to train for 2-way, 3-way, and 6-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 and contains text that is true (i.e. direct quotes from propaganda posters), or the sample is fake with false text. Our final 6-way classification was created to categorize different types
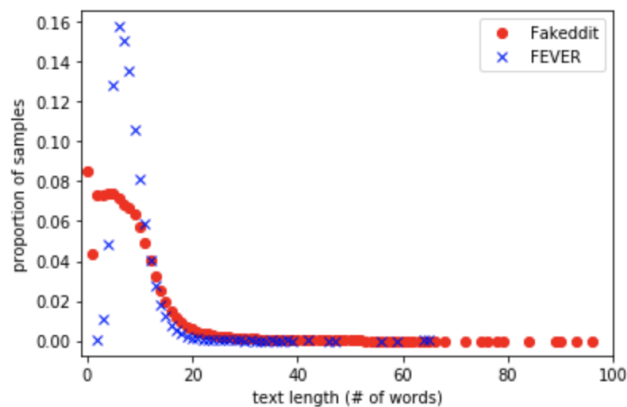
Figure 2: Distributions of word length in Fakeddit and FEVER datasets. We exclude samples that have more than 100 words.
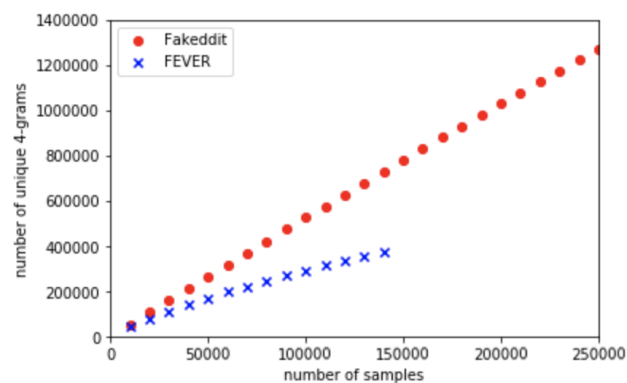

Figure 3: Type-caption curve of Fakeddit vs. FEVER with 4-gram type.

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. In addition, it will enable researchers to focus on specific types of fake news classification if they desire; for example, focusing on satire only. For the 6-way classification, the first label is true and the other five are defined within the seven types of fake news (Wardle, 2017). Only five types of fake news were chosen as we did not find subreddits with posts aligning with the remaining two types. We provide examples from each class for 6-way classification in Figure 1. The 6-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 and mildlyinteresting. The former consists of posts from various news sites. The latter encompasses real photos with accurate captions.

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, with headlines such as "Man Lowers Carbon Footprint By Bringing Reusable Bags Every Time He Buys Gas".

Misleading Content: This category consists of informa-

Dataset	1-gram	2-gram	3-gram	4-gram
FEVER	40874	179525	315025	387093
Fakeddit	61141	507512	767281	755929

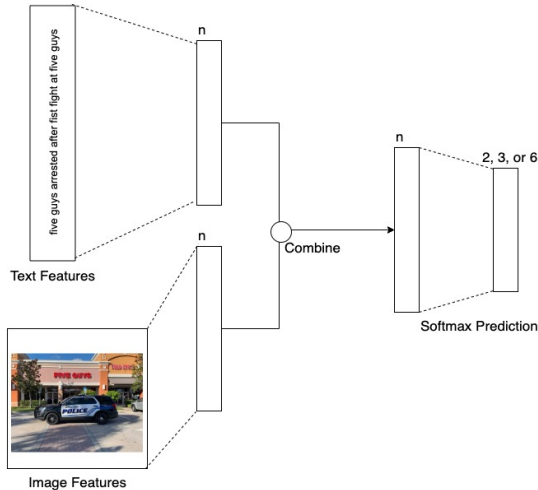Table 3: Unique n-grams for FEVER and Fakeddit for equal sample size (FEVER's total dataset size).

Figure 4: Multimodal model for integrating text and image data for 2, 3, and 6-way classification. n, the hidden layer size, is tuned for each model instance through hyperparameter optimization.

tion that is intentionally manipulated to fool the audience. Our dataset contains three subreddits in this category.

Imposter Content: This category contains two subreddits, which contain bot-generated content and are trained on a large number of other subreddits.

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 that do not relate to the true meaning of the image.

Manipulated Content: Content that has been purposely manipulated through manual photo editing or other forms of alteration. The photoshopbattle subreddit comments (not submissions) make up the entirety of this category. Samples contain doctored derivatives of images from the submissions.

3.4. Dataset Analysis

In Table 2, we provide an overview of specific statistics pertaining to our dataset such as vocabulary size and number of unique users. We also provide a more in-depth analysis in comparison to another sizable dataset, FEVER.

First, we choose to examine the word lengths of our text data. Figure 2 shows the proportion of samples per text length for both Fakeddit and FEVER. It can be seen that our dataset contains a higher proportion of longer text starting from word lengths of around 17, while FEVER's captions peak at around 10 words. In addition, while FEVER's peak is very sharp, Fakeddit has a much smaller and more gradual slope. Fakeddit also provides a broader diversity of text lengths, with samples containing almost 100 words.

Meanwhile, FEVER's longest text length stops at less than 70 words.

Secondly, we examine the linguistic variety of our dataset by computing the Type-Caption Curve, as defined in (Wang et al., 2019). Figure 3 shows these results. Fakeddit provides significantly more lexical diversity. Even though Fakeddit contains more samples than FEVER, the number of unique n-grams contained in similar sized samples are still much higher than those within FEVER. These effects will be magnified as Fakeddit contains more than 5 times more total samples than FEVER. In Table 3, we show the number of unique n-grams for both datasets when sampling n samples, where n is equal to FEVER's dataset size. This demonstrates that for all n-gram sizes, our dataset is more lexically diverse than FEVER's for equal sample sizes.

These salient text features - longer text lengths, broad array of text lengths, and higher linguistic variety - highlight Fakeddit's diversity. This diversity can strengthen fake news detection systems by increasing their lexical scope.

4. Experiments

4.1. Fake News Detection

Multiple methods were employed for text and image feature extraction. We used InferSent (Conneau et al., 2017) and BERT (Devlin et al., 2019) to generate text embeddings for the title of the Reddit submissions. VGG16 (Simonyan and Zisserman, 2015), EfficientNet (Tan and Le, 2019), and ResNet50 (He et al., 2016) were utilized to extract the features of the Reddit submission thumbnails.

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 embeddings (Joulin et al., 2017). We obtained encoded sentence features of length 4096 for each submission title using InferSent.

In addition, we used the BERT model. BERT 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(Xiao, 2018) tool, to map variable-length text/sentences into a 768 element array for each Reddit submission title. For our experiments, we utilized the pre-trained BERT-Large, Uncased model.

We employed 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 variation: B4. This was chosen as it is comparable to ResNet50 in terms of FLOP count. For the image models, we preloaded weights of models trained on ImageNet and included the top layer and used the penultimate layer for feature extraction.

4.2. Experiment Settings

As mentioned in section 3.2, the text was cleaned thoroughly through a series of steps. We also prepared the images by constraining the sizes of the images to match the input size of the image models. We applied necessary image preprocessing required for the image models.

			2-way		3-way		6-way	
Type	Text	Image	Validation	Test	Validation	Test	Validation	Test
Text	BERT	–	**0.8654**	**0.8644**	**0.8582**	**0.8580**	**0.7696**	**0.7677**
	InferSent	–	0.8634	0.8631	0.8569	0.8570	0.7652	0.7666
Image	–	VGG16	0.7355	0.7376	0.7264	0.7293	0.6462	0.6516
	–	EfficientNet	0.6115	0.6087	0.5877	0.5828	0.4152	0.4153
	–	ResNet50	**0.8043**	**0.8070**	**0.7966**	**0.7988**	**0.7529**	**0.7549**
Text+Image	InferSent	VGG16	0.8655	0.8658	0.8618	0.8624	0.8130	0.8130
	InferSent	EfficientNet	0.8328	0.8339	0.8259	0.8256	0.7266	0.7280
	InferSent	ResNet50	0.8888	0.8891	0.8855	0.8863	0.8546	0.8526
	BERT	VGG16	0.8694	0.8699	0.8644	0.8655	0.8177	0.8208
	BERT	EfficientNet	0.8334	0.8318	0.8265	0.8255	0.7258	0.7272
	BERT	ResNet50	**0.8929**	**0.8909**	**0.8905**	**0.8890**	**0.8600**	**0.8588**

Table 4: Results on fake news detection for 2, 3, and 6-way classification with combination method of maximum.

	2-way		3-way		6-way	
Combination Methods	Validation	Test	Validation	Test	Validation	Test
Add	0.8551	0.8551	0.8509	0.8505	0.8206	0.8235
Concatenate	0.8564	0.8568	0.8531	0.8530	0.8237	0.8249
Maximum	**0.8929**	**0.8909**	**0.8905**	**0.8890**	**0.8600**	**0.8588**
Average	0.8554	0.8561	0.8512	0.8518	0.8229	0.8242

Table 5: Results on different multi-modal combinations for BERT + ResNet50

For our experiments, we excluded submissions that have either text or image data missing. We performed 2-way, 3-way, and 6-way classification for each of the three types of inputs: image only, text only, and multimodal (text and image). As in Figure 4, when combining the features in multimodal classification, we first condensed them into n-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. For all experiments, we tuned the hyperparameters on the validation dataset using the keras-tuner tool[10]. Specifically, we employed the Hyperband tuner (Li and Jamieson, 2018) to find optimal hyperparameters for the hidden layer size and learning rates. The hyperparameters are tuned on the validation set. We varied the number of units in the hidden layer from 32 to 256 with increments of 32. For the optimizer, we used Adam (Kingma and Ba, 2014) and tested three learning rate values: 1e-2, 1e-3, 1e-4. For the multimodal model, the unit size hyperparameter affected the sizes of the 3 layers simultaneously: the 2 layers that are combined and the layer that is the result of the combination. For non-multimodal models, we utilized a single size-tunable hidden layer, followed by a softmax predictor. For each model, we specified a maximum of 20 epochs and an early stopping callback to halt training if the validation accuracy decreased.

4.3. Results

The results are shown in Tables 4 and 5. For image and multimodal classification, ResNet50 performed the best followed by VGG16 and EfficientNet. In addition, BERT achieved better results than InferSent for multimodal classification. Multimodal features performed the best, followed

---

[10]https://github.com/keras-team/keras-tuner

by text-only, and image-only. Thus, image and text multi-modality present in our dataset significantly improves fake news detection. The "maximum" method to merge image and text features yielded the highest accuracy. Overall, the multimodal model that combined BERT text features and ResNet50 image features through the maximum method performed most optimally. The best 6-way classification model parameters were: hidden layer sizes of 224 units, 1e-4 learning rate, trained over 20 epochs.

5. Error Analysis

We conduct an error analysis on our 6-way detection model by examining samples from the test set that the model predicted incorrectly. A subset of these samples is shown in Table 6. Firstly, the model had the most difficult time identifying imposter content. This category contains subreddits that contain machine-generated samples. Recent advances in machine learning such as Grover (Zellers et al., 2019), a model that produces realistic-looking machine-generated news articles, have allowed machines to automatically generation human-like material. Our model has a relatively difficult time identifying these samples. The second category the model had the poorest performance on was satire samples. The model may have a difficult time identifying satire because creators of satire tend to focus on creating content that seems similar to real news if one does not have a sufficient level of contextual knowledge. Classifying the data into these two categories (imposter content and satire) are complex challenges, and our baseline results show that there is significant room for improvement in these areas. On the other hand, the model was able to correctly classify almost all manipulated content samples. We also found that misclassified samples were frequently categorized as being true. This can be attributed to the relative size of true samples in the 6-way classification. While we have compara-

Text	Image	Predicted Label	Gold Label	PM(%)
volcanic eruption in bali last night		False Connection	True	17.9
nascar race stops to wait for family of ducks to pass		True	Satire	32.8
cars race towards nuclear explosion		True	False Connection	17.8
bear experiences getting hit in the cinema rule, your child again		Satire	Imposter Content	55.7
three corgis larping at the beach		True	Manipulated Content	3.3
mighty britain getting tied down in south africa during boer bar circa		False Connection	Misleading Content	16.9

Table 6: Classification errors on the BERT+ResNet50 model for 6-way classification. PM: Proportion of samples misclassified within each Gold label.

ble sizes of fake and true samples for 2-way classification, 6-way breaks down the fake news into more fine-grained classes. As a result, the model trains on a higher number of true samples and may be inclined to predict this label.

conducted using our dataset's unique multimodality aspect. We hope that our dataset can be used to advance efforts to combat the ever-growing rampant spread of disinformation in today's society.

6. Conclusion

In this paper, we presented a novel dataset for fake news research, Fakeddit. Compared to previous datasets, Fakeddit provides a large number of multimodal samples with multiple labels for various levels of fine-grained classification. We conducted several experiments with multiple baseline models and performed an error analysis on our results, highlighting the importance of large scale multimodality unique to Fakeddit and demonstrating that there is still significant room for improvement in fine-grained fake news detection. Our dataset has wide-ranging practicalities in fake news research and other research areas. Although we do not utilize submission metadata and comments made by users on the submissions, we anticipate that these additional multimodal features will be useful for further fake news research. For example, future research can look into tracking a user's credibility through using the metadata and comment data provided and incorporating video data as another multimedia source. Implicit fact-checking research with an emphasis on image-caption verification can also be

Acknowledgments

7. Bibliographical References

Abu Salem, F. K., Al Feel, R., Elbassuoni, S., Jaber, M., and Farah, M. (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, Jul.

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

Boididou, C., Papadopoulos, S., Zampoglou, M., Apostolidis, L., Papadopoulou, O., and Kompatsiaris, Y. (2018). Detection and visualization of misleading con-

Subreddit	6-Way Label	URL
photoshopbattles submissions	True	https://www.reddit.com/r/photoshopbattles
nottheonion	True	https://www.reddit.com/r/nottheonion
neutralnews	True	https://www.reddit.com/r/neutralnews
pic	True	https://www.reddit.com/r/pic
usanews	True	https://www.reddit.com/r/usanews
upliftingnews	True	https://www.reddit.com/r/upliftingnews
mildlyinteresting	True	https://www.reddit.com/r/mildlyinteresting
usnews	True	https://www.reddit.com/r/usnews
fakealbumcovers	Satire	https://www.reddit.com/r/fakealbumcovers
satire	Satire	https://www.reddit.com/r/satire
waterfordwhispersnews	Satire	https://www.reddit.com/r/waterfordwhispersnews
theonion	Satire	https://www.reddit.com/r/theonion
propagandaposters	Misleading Content	https://www.reddit.com/r/propagandaposters
fakefacts	Misleading Content	https://www.reddit.com/r/fakefacts
savedyouaclick	Misleading Content	https://www.reddit.com/r/savedyouaclick
misleadingthumbnails	False Connection	https://www.reddit.com/r/misleadingthumbnails
confusing_perspective	False Connection	https://www.reddit.com/r/confusing_perspective
pareidolia	False Connection	https://www.reddit.com/r/pareidolia
fakehistoryporn	False Connection	https://www.reddit.com/r/fakehistoryporn
subredditsimulator	Imposter Content	https://www.reddit.com/r/subredditsimulator
subsimulatorgpt2	Imposter Content	https://www.reddit.com/r/subsimulatorgpt2
photoshopbattles comments	Manipulated Content	https://www.reddit.com/r/photoshopbattles

Table 7: List of subreddits used in Fakeddit.

tent on twitter. *International Journal of Multimedia Information Retrieval*, 7(1):71–86.

Christlein, V., Riess, C., Jordan, J., Riess, C., and Angelopoulou, E. (2012). An evaluation of popular copy-move forgery detection approaches. *IEEE Transactions on information forensics and security*, 7(6):1841–1854.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. (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, September. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Dreyfuss, E. and Lapowsky, I. (2019). Facebook is changing news feed (again) to stop fake news. *Wired*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Heller, S., Rossetto, L., and Schuldt, H. (2018). The PS-Battles Dataset – an Image Collection for Image Manipulation Detection. *CoRR*, abs/1804.04866.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2017). Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain, April. Association for Computational Linguistics.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, L. and Jamieson, K. (2018). Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18:1–52.

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

Nørregaard, J., Horne, B. D., and Adali, S. (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, Jul.

Pathak, A. and Srihari, R. (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, July. Association for Computational Linguistics.

Santia, G. C. and Williams, J. R. (2018). Buzzface: A news veracity dataset with facebook user commentary and egos. In *Twelfth International AAAI Conference on Web and Social Media*.

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., and Liu, H. (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.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

Tacchini, E., Ballarin, G., Vedova, M. L. D., Moret, S., and de Alfaro, L. (2017). Some like it hoax: Automated fake news detection in social networks. *CoRR*, abs/1704.07506.

Tan, M. and Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks.

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. (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, June. Association for Computational Linguistics.

Wang, X., Wu, J., Chen, J., Li, L., Wang, Y.-F., and Wang, W. Y. (2019). Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *The IEEE International Conference on Computer Vision (ICCV)*, October.

Wang, W. Y. (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, July. Association for Computational Linguistics.

Wardle, C. (2017). Fake news. it's complicated. *First Draft*.

Xiao, H. (2018). bert-as-service. `https://github.com/hanxiao/bert-as-service`.

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., and Choi, Y. (2019). Defending against neural fake news. In *Advances in Neural Information Processing Systems 32*.

Zlatkova, D., Nakov, P., and Koychev, I. (2019). Fact-checking meets fauxtography: Verifying claims about images. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2099–2108, Hong Kong, China, November. Association for Computational Linguistics.

Zubiaga, A., Liakata, M., Procter, R., Hoi, G. W. S., and Tolmie, P. (2016). Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 11(3):e0150989.

Appendix

We show the list of subreddits in Table 7.