Volta at SemEval-2021 Task 6: Towards Detecting Persuasive Texts and Images using Textual and Multimodal Ensemble

Kshitij Gupta[*]   **Devansh Gautam**[*]   **Radhika Mamidi**

International Institute of Information Technology Hyderabad

{kshitij.gupta,devansh.gautam}@research.iiit.ac.in,
radhika.mamidi@iiit.ac.in

Abstract

Memes are one of the most popular types of content used to spread information online. They can influence a large number of people through rhetorical and psychological techniques. The task, *Detection of Persuasion Techniques in Texts and Images*, is to detect these persuasive techniques in memes. It consists of three subtasks: (A) Multi-label classification using textual content, (B) Multi-label classification and span identification using textual content, and (C) Multi-label classification using visual and textual content. In this paper, we propose a transfer learning approach to fine-tune BERT-based models in different modalities. We also explore the effectiveness of ensembles of models trained in different modalities. We achieve an F1-score of 57.0, 48.2, and 52.1 in the corresponding subtasks.

1 Introduction

Memes are text superimposed on graphics that convey messages in the form of jokes, sarcasm, etc. In the current era of the internet and social media, they are very quick to spread. If used as a part of a disinformation campaign, it can be quite tricky to notice the agenda behind them and has the potential to influence a large mass of people without them realizing it (Muller, 2018; Tardáguila et al., 2018; Glowacki et al., 2018).

To this end, SemEval 2021 Task 6 (Dimitrov et al., 2021) focuses on identifying such persuasive techniques (Miller, 1939) in a multimodal (visual-linguistic) setting. It consists of three subtasks that enable the participants to study the problem in each modality. Only the English textual cues are used in tasks A and B, while the visual cues are also used in task C. Task B is a modification of task A which further requires predicting the spans for each identified technique as well.

Figure 1: Sample memes demonstrating the multimodal setting

Meme classification is a multimodal problem that often requires visual and textual cues to convey a message. Memes can convey very different meanings if either of the cues is removed. A few samples are shown in figure 1 to demonstrate the importance of visual and textual cues for classification.

We experiment with BERT (Devlin et al., 2019) based unimodal models for tasks A and B. Since they are state-of-the-art models for natural language understanding, they are a good choice for understanding the complex propaganda techniques in texts. Transformers (Vaswani et al., 2017) have limited sequence length, which limits their performance on longer data, but in the case of memes, the textual content is also very limited.

For task C, we experiment with Visual-Linguistic (VL) models like UNITER (Chen et al., 2020), VisualBERT (Li et al., 2019), LXMERT (Tan and Bansal, 2019) for the cross-modal understanding of memes. We further explore the effectiveness of ensembling models trained in

---

[*]The authors have contributed equally.

different modalities.

The code for all subtasks is available at `http://github.com/kshitij98/multimodal-propaganda.git`

2 Background

Propaganda aims to push biased agendas to influence people's mindsets. It is successful in achieving its goal by hiding its way through any of the numerous media platforms available in the current world. A major factor behind the success of such campaigns is the boom of the internet and social media in recent years. Another factor being the difficulty to spot such techniques manually because of the high volume of text produced and the unnoticeable nature of such content. With the recent interest of the research community in "fake news," the detection of persuasive techniques or highly biased texts has emerged as an active research area. Some of the previous work in this direction analyzes the general pattern of propaganda (Garimella et al., 2018; Chatfield et al., 2015), performs analysis at a document level (Rashkin et al., 2017; Barrón-Cedeño et al., 2019) and a fine-grained analysis of the text (Da San Martino et al., 2019, 2020). However, most previous work analyses the techniques in a textual unimodal setting only. This work studies propaganda techniques in a new age domain like memes.

Meme classification task can be considered a combined VL multimodal problem. It is similar to the currently popular VL problems like Visual Question Answering (Antol et al., 2015), Visual Commonsense Reasoning (Zellers et al., 2019) and Visual Entailment (Xie et al., 2019), as we have to classify semantically correlated text with that of the visual content in the image. Hence a cross-modal approach under vision and text should perform better than unimodal architectures. Basic VL approaches are based on simple fusion techniques in the form of early or late fusion to correlate unimodally trained visual and textual models. However, in an ideal scenario, a multimodally trained model should be more effective in detecting persuasive techniques in memes. With the rising interest in VL problems, recent work attempts to study similar problems in a VL multimodal setting (Gomez et al., 2020; Kiela et al., 2020; Suryawanshi et al., 2020).

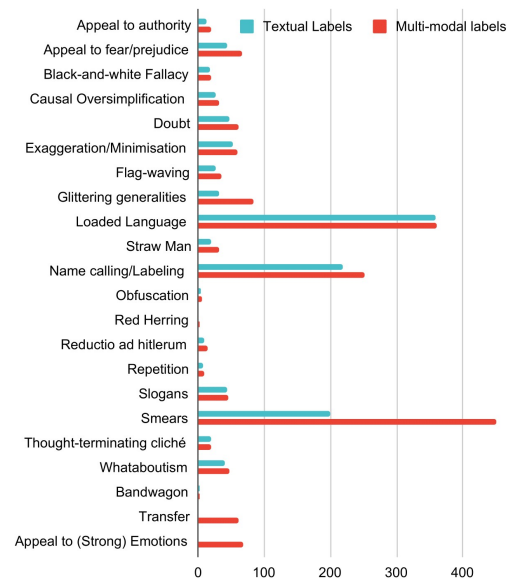**Data Description**   The dataset consists of 951 memes in total, which is further divided into train/

Figure 2: Data Distribution of the labels in the training set

dev/ test splits. All the tasks have the same set of memes in their training sets, but the labeling differs for each of them. For task A, only the textual cues were used to identify the techniques. For task B, the spans of each technique were further detected. For task C, more techniques were identified using visual cues.

The distribution of the labels is illustrated in figure 2. Detailed information of the dataset can be found in the task description paper (Dimitrov et al., 2021).

3 System Overview

Systems proposed for all the tasks use BERT-based models with task-specific modifications.

The proposed systems are explained in detail below:

3.1 Task A - Text Classification

We use transfer learning to fine-tune BERT-based models for this task. To fine-tune the models, we experiment with BASE and LARGE variants of BERT and RoBERTa. We attach a feed-forward network on the `[CLS]` token embedding with two linear layers having the model's default dropout and $Tanh$ activation layer in between. The model architecture is illustrated in figure 3a. We apply a standard Binary Cross Entropy loss to train this model with the hyperparameters mentioned in table 1. Since there is a substantial class imbalance in the dataset, we add weights to the positive samples in
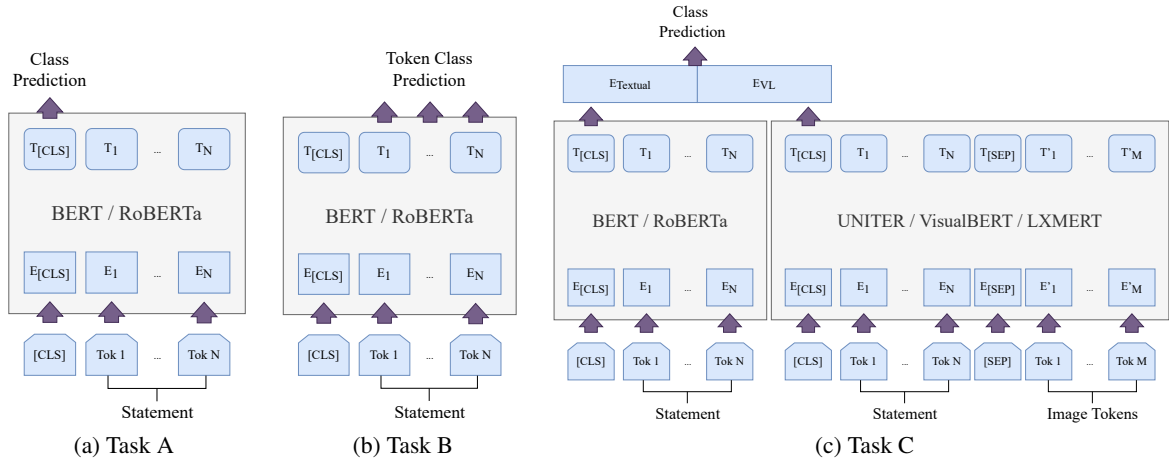
Figure 3: Proposed architectures for the given subtasks

the loss function by using the following equation:

$$\ell(\mathbf{x}, \mathbf{y}) = -\frac{1}{Nd} \sum_{n=1}^{N} \sum_{k=1}^{d} [p^k y_n^k \log x_n^k + (1 - y_n^k) \log(1 - x_n^k)]$$

$$p^k = \frac{1}{f^k}(|K| - f^k)$$

(1)

Where N is the batch size, n index denotes n^{th} batch element, d is the number of classes, f stands for a vector of class absolute frequencies calculated on the train set, $\mathbf{x}$ is the output vector from the last Sigmoid layer, $\mathbf{y}$ is a vector of multi-hot encoded ground truth labels and $|K|$ is the size of the train set.

We finally use the standard Sigmoid activation function to compute probabilities for each label.

3.2 Task B - Span Identification and Text Classification

We experiment with the same BERT-based encoders for this task.

Pre Processing Since the spans are given on character-level in the dataset and the transformer models run on token-level, we transform all the spans to token-level by taking the intersection with the tokenized input. Further, we train the model with the obtained token-level labels as targets.

Model We model the problem statement as a token-level *multi-label* classification problem for span identification. The model architecture is illustrated in figure 3b. We use the same classifier to classify all token embeddings. To handle class imbalance, we apply weighted Binary Cross-Entropy loss with class weights as mentioned in equation 1

Post Processing Since all tokens of a word should belong to the same set of classes, we merge all the tokens of each word and assign a union of those classes to the chosen word.

Finally, all the words are classified using the following equation where W_i is the set of labels of the ith word, $L_{i,j}$ is the set of labels of the jth token of the ith word and N_i is the number of tokens of the ith word:

$$W_i = \bigcup_{j=1}^{N_i} L_{i,j}$$

We do not apply loss on special tokens so that the classifier is not misled while training.

3.3 Task C - Visual-Linguistic Classification

We experiment with Visual-Linguistic (VL) models for this task. Since meme classification is a multimodal task, a multimodal transformer architecture should be a good choice.

Pre Processing The image tokens used in the VL models are extracted using Faster R-CNN (Ren et al., 2015) and a fixed number of image tokens are created by extracting the features of the top 36 regions of interest.

Model We first experiment with training the VL models directly. Although the VL models should perform better in the presence of extra cues for classification, poor performance is observed compared to the powerful textual-only models.

On further analyzing the problem, we observe that propaganda detection is a complex semantic problem, and often the classes can be detected by

1077

Parameter	Task A	Task B	Task C	
			VL	Ensemble
Dropout	0.1	-	-	-
Max Sequence Length	128	512	128	128
Batch Size	8	8	16	32
warmup	-	-	0.1	0.1
Learning Rate	1e-05	1e-05	1e-05	1e-05
patience	50	50	50	50
Weight Decay	-	0.01	0.01	0.01
Optimizer	Adam	AdamW	BertAdam	BertAdam

Table 1: Hyperparameters

using just the text. So, we experiment with the textual models and ignore the visual cues in the data. Surprisingly, the textual model outperforms the VL model when trained to learn multimodal labels using just the textual cues. We further study both models to better understand the learnings of each of them (see Section 5) and continue experimentation with ensembling both models.

We propose an ensemble of multimodal transformers like UNITER, VisualBERT, LXMERT with unimodal transformers like BERT and RoBERTa to help the classification model as each architecture can focus on their domains and later merge those embeddings. Rather than using a naive average ensembling method, we propose our own model, illustrated in figure 3c. Our model concatenates the [CLS] token embeddings from transformers in each modality and then applies classification on top of the concatenated vector. The base encoders are fine-tuned unimodally and then frozen while training the ensemble classifier. To train the textual model unimodally, we train the textual model with textual labels only to learn relevant features.

4 Experimental Setup

4.1 Implementation

We use HuggingFace [1] library (Wolf et al., 2020) to experiment with textual models, and use the official implementations of the multimodal transformers in PyTorch[2] (Paszke et al., 2019). The models were trained with the default hyperparameters with an exception for the parameters mentioned in table 1. All experiments were conducted using Nvidia GeForce RTX 2080 Ti GPU.

For training the models, we use the same train/ dev/ test split as mentioned in the task, which is

in the ratio 10:1:3 with 951 samples in total. We use the best-performing model by comparing the F1-micro scores on the validation set for all our experiments.

4.2 Evaluation Metrics

For tasks A and C, the F1-micro score is used as the main performance metric. Due to various low resource classes in the dataset, F1-macro was also used to give weight to the smaller classes in the performance metric, but it was revealed to be highly impacted by small variations in the model.

For task B, we use the official evaluation metric as defined in the task. As the task is a multi-label sequence tagging task, the standard micro-averaged F1 is modified to account for partial matching between the spans.

5 Results

We conduct several experiments to compare the performance of models on each of the tasks. Detailed information can be seen in table 2.

For task A, RoBERTa $_{LARGE}$ is the best performing model on the task. Although the average performance of the model is worse than BERT $_{LARGE}$ on the test set, the maximum performance is still better.

Similarly, for task B, RoBERTa $_{LARGE}$ outperforms other models by a large margin.

For task C, we trained the textual models on visual-linguistic labels for a fair comparison with the other VL models. Unexpectedly, they also show decent performance and surprisingly perform better than the multimodal transformers. Finally, the ensemble models are trained with their textual model trained on textual labels and the VL model trained on VL labels. They outperform their unimodal components and end up with a fair performance increase in each of the cases.

During the post-evaluation phase, we conduct more experiments and report the corresponding performances on the test set for the best-validated models on the dev set. We realize that our submitted model performed worse on the test set in task C because the checkpoint used for test set predictions was different from the best validation checkpoint.

The final values we achieve on the test set by using the best validation set checkpoints are 57.56, 48.23, and 55.68 by using RoBERTa $_{LARGE}$, RoBERTa $_{LARGE}$ and RoBERTa $_{LARGE}$ + UNITER, respectively on all the subtasks.

---

[1] 🤗 Transformers, v4.2.0, https://huggingface.co/transformers/

[2] PyTorch, v1.7.1, https://pytorch.org/

Model	Task A		Task B		Task C	
	Dev Set	Test Set	Dev Set	Test Set	Dev Set	Test Set
BERT $_{\text{BASE}}$	$64_{\pm 0.7}$	$52.2_{\pm 2.5}$	$53.1_{\pm 1}$	$45.6_{\pm 1.5}$	$63.3_{\pm 1.2}$	$51.9_{\pm 1}$
BERT $_{\text{LARGE}}$	$62.8_{\pm 0.8}$	$\mathbf{54.8}_{\pm 1.5}$	$53.5_{\pm 2.1}$	$44.7_{\pm 2.5}$	$62.7_{\pm 1.4}$	$52.5_{\pm 0.9}$
RoBERTa $_{\text{BASE}}$	$61_{\pm 0.8}$	$51.2_{\pm 1.2}$	$53.2_{\pm 0.6}$	$43.9_{\pm 0.9}$	$61.5_{\pm 1.1}$	$49.8_{\pm 1.2}$
RoBERTa $_{\text{LARGE}}$	$\mathbf{64.7}_{\pm 1.1}$	$53.2_{\pm 3.9}$	$\mathbf{58.5}_{\pm 2.1}$	$\mathbf{47.6}_{\pm 1.5}$	$63.9_{\pm 1.1}$	$54.2_{\pm 1}$
UNITER					$64.9_{\pm 0.2}$	$49.2_{\pm 0.6}$
VisualBERT					$57.8_{\pm 0.6}$	$45.8_{\pm 0.7}$
LXMERT					$54.5_{\pm 0.3}$	$44.4_{\pm 0.0}$
BERT $_{\text{BASE}}$ + UNITER					$66.1_{\pm 0.2}$	$52.9_{\pm 0.8}$
RoBERTa $_{\text{LARGE}}$ + UNITER					$\mathbf{67.3}_{\pm 0.9}$	$\mathbf{54.9}_{\pm 0.6}$
BERT $_{\text{BASE}}$ + VisualBERT					$63.1_{\pm 0.3}$	$53.1_{\pm 0.5}$
RoBERTa $_{\text{LARGE}}$ + VisualBERT					$64.4_{\pm 0.5}$	$52.7_{\pm 1.9}$

Table 2: Mean and std of F1-micro scores computed from 10 runs of the mentioned models. Test set values are reported after choosing the best model for the dev set using early stopping.
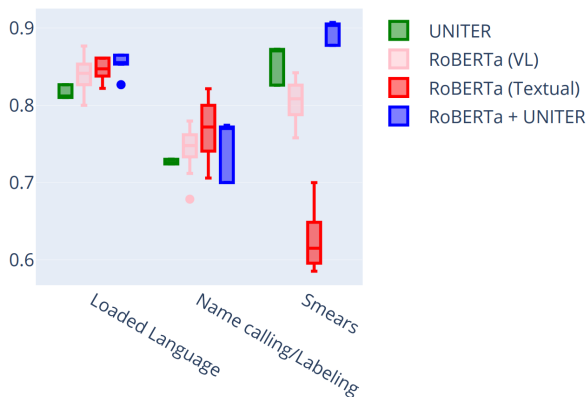

Figure 4: Comparison of different models in the 3 most occurring classes. F1-micro scores of 10 runs are computed on the dev set.

We observe that all BERT-based models give a decent performance for all tasks. An interesting insight we get from the visual-linguistic task is that the problem is not completely multimodal. Several of the persuasive techniques can be identified by using just the textual cues, which is evident from figure 2 as well. While textual models like RoBERTa are pretrained on a much larger textual dataset and are able to learn more complex semantics from the data, VL models suffer when used in textual-only domains. We carry out experiments to further study the differences in these models with different types of labels and propose an architecture that benefits from both models.

A class-wise comparison is carried out for all models. Since several of the classes have very low positive samples, it is difficult to draw conclusions because of the high variance in performance due to different model initializations, so we compare the most frequently occurring labels only. The difference in the performance of different models is illustrated in figure 4. For this comparison, we train the textual model on multimodal and textual labels to compare the capability of the models. Ideally, the textual models should be trained on textual labels only to learn more relevant features. The ensemble is trained with a textual model trained on textual labels and a VL model trained on VL labels. The comparison shows that the multimodal models are performing much better for detecting *Smears*, which has various samples which require visual cues as well. Another observation is that RoBERTa is performing better on *Loaded Language* and *Name calling / Labelling*. Training an ensemble of the textual model and the multimodal is helping the model perform better in all classes.

To further study the models, we report the performance for each of them in different modalities. To measure the textual performance of the models, we compare the sets of labels in task A and C to shortlist labels which were identified only after the presence of visual cues. We do not consider those predictions and calculate the F1-micro score for the remaining subset of the predictions. Similarly, for measuring the visual performance, we do not consider the labels which were identified by just using the textual cues. The performance of several runs of the models is compared in figure 5. The comparison also supports the claim that ensembling models trained in different modalities help to learn from
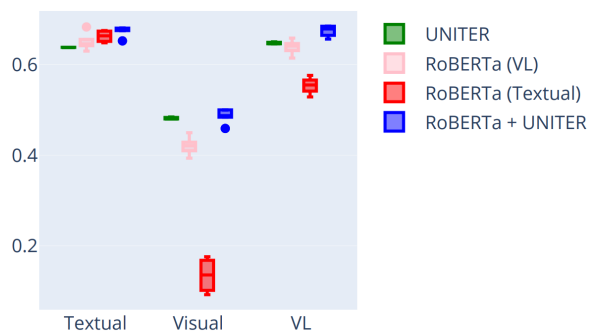
Figure 5: Comparison of different models in different modes. F1-micro scores of 10 runs are computed on the dev set.

the best of both worlds.

6 Conclusion

Although detecting persuasive techniques in memes is a multimodal problem, often, most of the techniques can be identified by just using the textual cues from the meme. Since VL models are still in their nascent stages, powerful textual models help with solving the problem at hand. Future work can be done to improve these kinds of problems that are not multimodal in the truest sense. The ensembling method used in our model is very basic; better architectures can be explored to continue this line of work.

Acknowledgments

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Alberto Barrón-Cedeño, Giovanni Da San Martino, Israa Jaradat, and Preslav Nakov. 2019. Proppy: A system to unmask propaganda in online news. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):9847–9848.

Akemi Takeoka Chatfield, Christopher G. Reddick, and Uuf Brajawidagda. 2015. Tweeting propaganda, radicalization and recruitment: Islamic state supporters multi-sided twitter networks. In *Proceedings of the 16th Annual International Conference on Digital Government Research*, dg.o '15, page 239–249, New York, NY, USA. Association for Computing Machinery.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *ECCV*.

Giovanni Da San Martino, Alberto Barrón-Cedeño, and Preslav Nakov. 2019. Findings of the NLP4IF-2019 shared task on fine-grained propaganda detection. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 162–170, Hong Kong, China. Association for Computational Linguistics.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online). International Committee for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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. Association for Computational Linguistics.

Dimiter Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. Task 6 at semeval-2021: Detection of persuasion techniques in texts and images. In *Proceedings of the 15th International Workshop on Semantic Evaluation*, SemEval '21, Bangkok, Thailand.

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Quantifying controversy on social media. *Trans. Soc. Comput.*, 1(1).

Monika Glowacki, Vidya Narayanan, Sam Maynard, Gustavo Hirsch, Bence Kollanyi, Lisa-Maria Neudert, Phil Howard, Thomas Lederer, and Vlad Barash. 2018. News and political information consumption in Mexico: Mapping the 2018 Mexican presidential election on Twitter and Facebook. Technical Report COMPROP DATA MEMO 2018.2, Oxford University, Oxford, UK.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and

Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. In *Arxiv*.

Clyde R. Miller. 1939. The Techniques of Propaganda. From "How to Detect and Analyze Propaganda," an address given at Town Hall. The Center for learning.

Robert Muller. 2018. Indictment of Internet Research Agency. pages 1–37.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41, Marseille, France. European Language Resources Association (ELRA).

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Cristina Tardáguila, Fabrício Benevenuto, and Pablo Ortellado. 2018. Fake News Is Poisoning Brazilian Politics. WhatsApp Can Stop It. https://www.nytimes.com/2018/10/17/opinion/brazil-election-fake-news-whatsapp.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.