A Multi-Modal Method for Satire Detection using Textual and Visual Cues

Lily Li[1], **Or Levi**[2], **Pedram Hosseini**[3], **David A. Broniatowski**[3]
[1]Jericho Senior High School, New York, USA
[2]AdVerifai, Amsterdam, Netherlands
[3]The George Washington University, Washington D.C., USA
lily.li@jerichoapps.org, or@adverifai.com
{phosseini,broniatowski}@gwu.edu

Abstract

Satire is a form of humorous critique, but it is sometimes misinterpreted by readers as legitimate news, which can lead to harmful consequences. We observe that the images used in satirical news articles often contain absurd or ridiculous content and that image manipulation is used to create fictional scenarios. While previous work have studied text-based methods, in this work we propose a multi-modal approach based on state-of-the-art visiolinguistic model ViLBERT. To this end, we create a new dataset consisting of images and headlines of regular and satirical news for the task of satire detection. We fine-tune ViLBERT on the dataset and train a convolutional neural network that uses an image forensics technique. Evaluation on the dataset shows that our proposed multi-modal approach outperforms image-only, text-only, and simple fusion baselines.

1 Introduction

Satire is a literary device that writers employ to mock or ridicule a person, group, or ideology by passing judgment on them for a cultural transgression or poor social behavior. Satirical news utilizes humor and irony by placing the target of the criticism into a ridiculous, fictional situation that the reader must suspend their disbelief and go along with (Maslo, 2019). However, despite what absurd content satirical news may contain, it is often mistaken by readers as real, legitimate news, which may then lead to the unintentional spread of misinformation. In a recent survey conducted by The Conversation (Garrett et al., 2019), up to 28% of Republican respondents and 14% of Democratic respondents reported that they believed stories fabricated by the Babylon Bee, a satirical news website, to be "definitely true". In these instances, the consequences of satire are indistinguishable from those of fake news.

To reduce the spread of misinformation, social media platforms have partnered with third-party fact-checkers to flag false news articles and tag articles from known satirical websites as satire for users (Facebook, nd; Google, nd). However, due to the high cost and relative inefficiency of employing experts to manually annotate articles, many researchers have tackled the challenge of automated satire detection. Existing models for satirical news detection have yet to explore the visual domain of satire, even though image thumbnails of news articles may convey information that reveals or disproves the satirical nature of the articles. In the field of cognitive-linguistics, Maslo (2019) observed the use of altered images showing imaginary scenarios on the satirical news show The Daily Show. This phenomenon also extends to satirical news articles, as seen in Figure 1. For example, Figure 1(A) depicts the Marvel Cinematic Universe character Hulk from the film Avengers: Infinity War and the United States President Donald Trump spliced together. Alone, each of the two images is serious and not satirical, but, since they come from drastically different contexts, combining the two images creates a clearly ridiculous thumbnail that complements the headline of the article.

In our work, we propose a multi-modal method for detecting satirical news articles. We hypothesize that 1) the content of news thumbnail images when combined with text, and 2) detecting the presence of manipulated or added characters and objects, can aid in the identification of satirical articles.

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

Proceedings of the 3rd NLP4IF Workshop on NLP for Internet Freedom, pages 33–38
Barcelona, Spain (Online), December 12, 2020

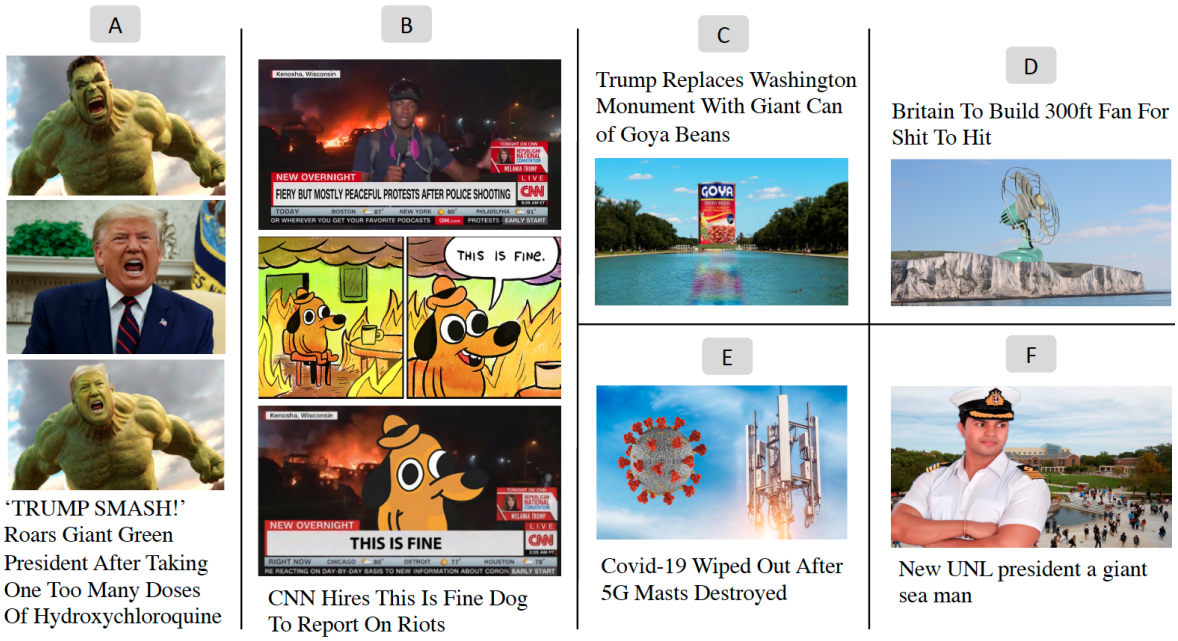

Figure 1: Examples of satirical news images created by altering existing images.

2 Related Work

Previous work proposed methods for satirical news detection using textual content (Levi et al., 2019). Some works utilize classical machine learning algorithms such as SVM with handcrafted features from factual and satirical news headlines and body text, including bag-of-words, n-grams, and lexical features (Burfoot and Baldwin, 2009; Rubin et al., 2016). More recent works use deep learning to extract learned features for satire detection. Yang et al. (2017) proposed a hierarchical model with attention mechanism and handcrafted linguistic features to understand satire at a paragraph and article-level.

While previous work utilize visiolinguistic data for similar tasks, there is no related work that employs multi-modal data to classify articles into satirical and factual news. Nakamura et al. (2019) created a dataset containing images and text for fake news detection in posts on the social media website Reddit. While they include a category for satire/parody in their 6-way dataset, since they use only content that has been submitted by Reddit users, it is not representative of mainstream news media. Multi-modal approaches have also been tried in sarcasm detection; Castro et al. (2019) compiled a dataset of scenes from popular TV shows and Cai et al. (2019) used tweets comprising of text and images from Twitter.

3 Methods

3.1 Data

We create a new multi-modal dataset of satirical and regular news articles. The satirical news is collected from four websites that explicitly declare themselves to be satire, and the regular news is collected from six mainstream news websites[1]. Specifically, the satirical news websites we collect articles from are The Babylon Bee, Clickhole, Waterford Whisper News, and The DailyER. The regular news websites are Reuters, The Hill, Politico, New York Post, Huffington Post, and Vice News. We collect the headlines and the thumbnail images of the latest 1000 articles for each of the publications. The dataset contains a total of 4000 satirical and 6000 regular news articles.

3.2 Proposed Models

Multi-Modal Learning. We use Vision & Language BERT (ViLBERT), a multi-modal model proposed by Lu et al. (2019) that processes images and text in two separate streams. Each stream consists of

[1]The regular news websites we use are listed by Media Bias/Fact Check `https://mediabiasfactcheck.com/`, a volunteer-run and nonpartisan organization dedicated to fact-checking and determining the bias of news publications

transformer blocks based on BERT (Devlin et al., 2018) and co-attentive layers that facilitate interaction between the visual and textual modalities. In each co-attentive transformer layer, multi-head attention is computed the same as a standard transformer block except the visual modality attends to the textual modality and vice-versa. To learn representations for vision-and-language tasks, ViLBERT is pre-trained using the masked multi-model modeling and multi-modal alignment prediction tasks on the Conceptual Captions dataset (Sharma et al., 2018). We choose to use ViLBERT because of its high performance on a variety of visiolinguistic tasks, including Visual Question Answering, Image Retrieval, and Visual Commonsense Reasoning. We fine-tune ViLBERT on the satire detection dataset by passing the element-wise product of the final image and text representations into a learned classification layer.

Image Forgery Detection. Since satirical news images are often forged from two or more images (known as image splicing), we implement an additional model that uses error level analysis (ELA). ELA is an image forensics technique that takes advantage of lossy JPEG compression for image tampering detection (Krawetz, 2007). In ELA, each JPEG image is resaved at a known compression rate, and the absolute pixel-by-pixel differences between the original and the resaved images are compared. ELA can be used to identify image manipulations where a lower quality image was spliced into a higher quality image or vice-versa. To detect image forgeries as an indicator of satirical news, we preprocess the images using ELA with a compression rate of 90% and use them as input into a CNN.

For the CNN, we use two convolutional layers with 32 kernels and a filter width of 5, each followed by a max-pooling layer. The output features from the CNN are fed into a MLP with a hidden size of 256 and a classification layer. We pretrain the model on the CASIA 2.0 image tampering detection dataset (Dong et al., 2013) before fine-tuning on the images of the satire detection dataset.

Implemention. We divide the data into training and test sets with a ratio of 80%:20%. We train all our models with a batch size of 32 and Adam optimizer. We use the MMF (Singh et al., 2020) implementation of ViLBERT and fine-tune it for 12 epochs with a learning rate of 5e-6. We extract Mask RCNN (He et al., 2017) features from the images in the dataset as visual input. The ViLBERT model has 6 transformer blocks in the visual stream and 12 transformer blocks in the textual stream. Our ELA+CNN model is trained with a learning rate of 1e-5 for 7 epochs.[2]

3.3 Baselines

To create fair baselines for our fine-tuned ViLBERT model, we train multi-modal models that use simple fusion. In the model denoted as Concatenation, ResNet-101 (He et al., 2016) and BERT features are concatenated and a MLP is trained on top. In the model denoted as Average fusion, the output of ResNet-101 and BERT are averaged. We choose these two models as our baselines to evaluate the effects of ViLBERT's early fusion of visual and textual representations and multi-modal pre-training on Conceptual Captions (Sharma et al., 2018). We also fine-tune uni-modal ResNet-101 and $\text{BERT}_{\text{BASE}}$ models to compare the performance of the multi-modal models to.

Type	Model	Accuracy	F1 score	AUC-ROC
	All regular news	60.00	—	50.00
Baselines	ResNet101	73.54	65.26	80.28
	$\text{BERT}_{\text{BASE}}$	91.33	88.64	96.77
	Simple fusion (average)	92.53	90.44	96.74
	Simple fusion (concatenation)	92.74	90.70	97.31
Proposed Models	ELA+CNN	44.20	51.86	44.61
	ViLBERT	**93.80**	**92.16**	**98.03**

Table 1: Model performance on satire detection dataset.

[2]Scripts for our experiments are available at: `https://github.com/lilyli2004/satire`

4 Results and Discussion

4.1 Experimental Results

We measure the performance of the proposed and baseline models using Accuracy, F1 score, and AUC-ROC metrics. The results are shown in Table 1. The models using only the visual modality (ResNet-101 and CNN+ELA) do not perform as well as the model that uses only the text modality ($\text{BERT}_{\text{BASE}}$). The simple fusion models (Average fusion, Concatenation) perform marginally better than $\text{BERT}_{\text{BASE}}$. ViLBERT outperforms the simple fusion multi-modal models because it uses early, deep fusion and has undergone multi-modal pre-training rather than only separate uni-modal visual and text pre-training. ViLBERT also performs almost 3.5 F1 points above the uni-modal $\text{BERT}_{\text{BASE}}$ model.

Surprisingly, the performance of the ELA+CNN model was very poor, achieving an accuracy worse than random chance. While this is not in line with our initial hypothesis, there might be several reasons for these results: Firstly, ELA is not able to detect image manipulations if the images have been resaved multiple times since after they have been compressed at a high rate there is little visible change in error levels (Krawetz, 2007). This makes it especially difficult to identify manipulation in images taken from the Internet, as they have usually undergone multiple resaves and are not camera originals. Additionally, although ELA can be used as a method to detect and localize the region of an image that has been potentially altered, it does not allow for the identification of what kind of image manipulation technique was used. This is important because even reputable news publications, such as Reuters and The Associated Press use Photoshop and other software to perform minor adjustments to photos, for example, to alter the coloring or lighting, or to blur the background (Schlesinger, 2007; The Associated Press, 2014).

Figure 2 shows examples from the satire detection dataset that illustrate the inconsistency of error level analysis in highlighting image manipulations. Both Figure 2(A) and Figure 2(B) are thumbnails from satirical articles that have clearly been fabricated. However, it is clear from the difference in ELA values that Figure 2(A) is a composite, while the ELA of Figure 2(B) is relatively uniform so the splicing can go undetected. Similarly, Figure 2(C) and Figure 2(D) are both thumbnails from factual articles, yet the drastic difference in ELA values of the building in Figure 2(C) indicates that it has undergone heavy editing while the ELA in Figure 2(D) does not.

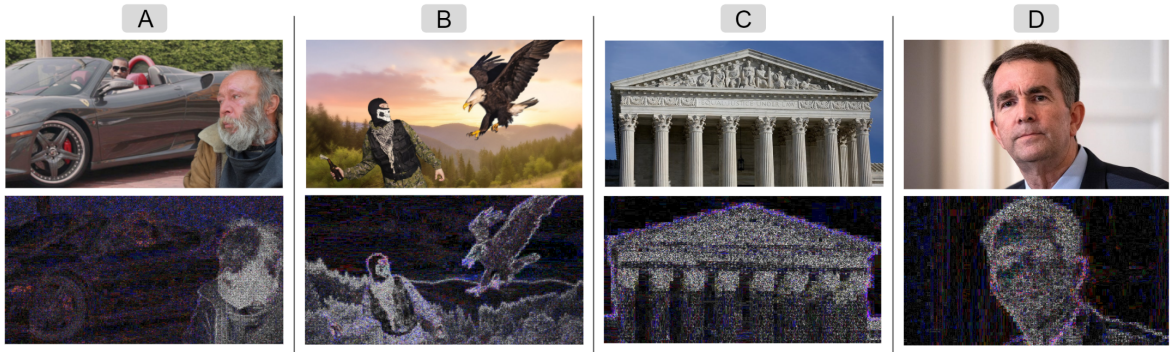

Figure 2: Examples of images in the satire detection dataset and their ELA.

4.2 Model Misclassification Study

After classification, we randomly select 20% of the test set samples misclassified by ViLBERT and observed them for patterns across multiple samples. Figure 3 shows examples of misclassified samples. We observed three main reasons that may have been the cause of the incorrectly classified articles: The model misinterpreted the headline (Figure 3(A)), the model lacks knowledge of current events (Figure 3(B)), and the article covered a bizarre but true story (Figure 3(C)).

Figure 3(A) shows an article from Politico that has been classified as satire. The image does not portray anything strange or out of the ordinary. However, the headline uses the word “bursts”, which the model might be incorrectly interpreting in the literal sense even though it is being used metaphorically.

If "bursts" was intended to be literal, it would drastically change the meaning of the text, which may be why the model failed to classify the article as factual. Figure 3(B) shows a satirical article from Babylon Bee that has been misclassified as factual. Its image has also not been heavily altered or faked; in fact, it is the same image that was used as the original thumbnail of the Joe Rogan podcast episode that is the subject of the article. However, the model fails to recognize the ridiculousness of the text, since it does not have the political knowledge to spot the contrast between the "alt-right" and the American politician Bernie Sanders. In Figure 3(C), an article is from the factual publication The New York Post is misclassified as satirical. Although both the headline and the image seem very ridiculous, the story and the image were, in fact, not fabricated. Thus, identifying text/images as absurd might not always aid in satire detection, since ViLBERT fails in classifying this article as factual because it is unable to tell that the image has not been forged.

Tammy Duckworth bursts into VP contention

Bernie Sanders Welcomed As Newest Member Of Alt-Right After Joe Rogan Endorsement

Cutout of Jeffrey Epstein appears in stands at UK soccer game

Figure 3: Examples of articles misclassified by ViLBERT

5 Conclusion and Future Investigations

In this paper, we create a multi-modal satire detection dataset and propose two models for the task based on the characteristics of satirical images and their relationships with the headlines. While our model based on image tampering detection performed significantly worse than the baselines, empirical evaluation showed the efficacy of our proposed multi-modal approach compared to simple fusion and uni-modal models. In future work on satire detection, we will incorporate image forensics methods to identify image splicing in satirical images, body text of articles instead of just headlines, as well as knowledge about politics and other current issues.

References

Clint Burfoot and Timothy Baldwin. 2009. Automatic satire detection: Are you having a laugh? In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 161–164, Suntec, Singapore, August. Association for Computational Linguistics.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2506–2515, Florence, Italy, July. Association for Computational Linguistics.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _Obviously_ perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629, Florence, Italy, July. 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.

Jing Dong, Wei Wang, and Tieniu Tan. 2013. CASIA image tampering detection evaluation database. In *2013 IEEE China Summit and International Conference on Signal and Information Processing*, pages 422–426. IEEE.

Facebook. n.d. *Fact-checking on Facebook: What publishers should know*. Business Help Center.

R. Kelly Garrett, Robert Bond, and Shannon Poulsen, 2019. *Too many people think satirical news is real*.

Google. n.d. *What does each label mean?* Publisher Help Center.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. 2017. Mask r-cnn. *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct.

Neal Krawetz. 2007. A picture's worth: Digital image analysis and forensics. *Black Hat Briefings*.

Or Levi, Pedram Hosseini, Mona Diab, and David A. Broniatowski. 2019. Identifying nuances in fake news vs. satire: Using semantic and linguistic cues. *arXiv preprint arXiv:1910.01160*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23.

Adi Maslo. 2019. Parsing satirical humor: a model of cognitive-linguistic satire analysis. *Književni jezik*, (30):231–253.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2019. r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection.

Victoria Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, pages 7–17, San Diego, California, June. Association for Computational Linguistics.

David Schlesinger. 2007. *The use of Photoshop*. Reuters Blogs Dashboard.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, July. Association for Computational Linguistics.

Amanpreet Singh, Vedanuj Goswami, Vivek Natarajan, Yu Jiang, Xinlei Chen, Meet Shah, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. 2020. Mmf: A multimodal framework for vision and language research. `https://github.com/facebookresearch/mmf`.

The Associated Press. 2014. *AP News Values and Principals*.

Fan Yang, Arjun Mukherjee, and Eduard Dragut. 2017. Satirical news detection and analysis using attention mechanism and linguistic features. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1979–1989, Copenhagen, Denmark, September. Association for Computational Linguistics.