



Delft University of Technology

N24News

A New Dataset for Multimodal News Classification

Wang, Zhen; Shan, Xu; Zhang, Xiangxie; Yang, Jie

Publication date

2022

Document Version

Final published version

Published in

2022 Language Resources and Evaluation Conference, LREC 2022

Citation (APA)

Wang, Z., Shan, X., Zhang, X., & Yang, J. (2022). N24News: A New Dataset for Multimodal News Classification. In N. Calzolari, F. Bechet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *2022 Language Resources and Evaluation Conference, LREC 2022* (pp. 6768-6775). (2022 Language Resources and Evaluation Conference, LREC 2022). European Language Resources Association (ELRA).

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

N24News: A New Dataset for Multimodal News Classification

Zhen Wang^{1*}, Xu Shan^{2*}, Xiangxie Zhang³, Jie Yang⁴

Delft University of Technology, Netherlands

{¹z.wang-42, ³x.zhang-60}@student.tudelft.nl, {²x.shan-2, ⁴j.yang-3}@tudelft.nl

Abstract

Current news datasets merely focus on text features on the news and rarely leverage the feature of images, excluding numerous essential features for news classification. In this paper, we propose a new dataset, *N24News*, which is generated from New York Times with 24 categories and contains both text and image information in each news. We use a multitask multimodal method and the experimental results show multimodal news classification performs better than text-only news classification. Depending on the length of the text, the classification accuracy can be increased by up to 8.11%. Our research reveals the relationship between the performance of a multimodal classifier and its sub-classifiers, and also the possible improvements when applying multimodal in news classification. *N24News* is shown to have great potential to prompt the multimodal news studies.

Keywords: Multimodal Dataset, News Article, Text Classification

1. Introduction

People have tried to use different carriers to record news. Ancient people first drew images by hands-on walls to record important things. After language was invented, words became the main tools for recording. Thanks to parchment preserved to this day, we can study people who lived a long time ago. Later, with the invention of the camera, images are widely used in news. Compared with text, images can bring us more intuitive information, even if we cannot understand the language used in the news. It is safe to say that images and text play an equally important role in news.

News classification is one of the essential tasks in news research (Katari and Myneni, 2020). We use the information provided by the news to group them into different categories. There is already some research about news classification, for example, news datasets, such as 20NEWS (Lang, 1995) and AG News (Zhang et al., 2015). However, they choose to ignore the images and merely pay attention to the text. This is not in line with the actual situation, especially when almost all the news today has images. In this work, we aim to use both images and text to achieve better news classification.

In order to combine heterogeneous information extracted from images and texts, multimodal methods are needed. Multimodal approaches can process various types of information simultaneously and has been used in news studies before. For example, in the fake news dataset Fakeddit (Nakamura et al., 2019), the authors propose a hybrid text+image model to classifier fake news. However, to best of our knowledge, currently there is no valid public news dataset containing enough real news with both images and texts that can be used to do multimodal news classification. Thus, in this work, we use the *New York Times* to build a new dataset called *N24News*. *N24News* is a large-scale multimodal news dataset comprising 60K image-text

pairs and 24 categories, which makes it possible to do multimodal real news classification tasks. Further, we use a multitask multimodal network to conduct a preliminary experiment in multimodal news classification, and the experiment shows the multimodal method can achieve higher accuracy than text-only news classification. Our error analysis reveals the relationship between the performance of a multimodal classifier and its sub-classifiers, and also the possible improvements when applying multimodal approaches in news classification.

2. Related Work

In news studies, the most commonly used datasets are 20NEWS (Lang, 1995) and AG NEWS (Zhang et al., 2015). 20NEWS is a collection of approximately 20,000 newsgroup documents across 20 different newsgroups, and AG News contains 1 million news articles gathered from more than 2000 news sources and grouped into four categories. These two datasets are now used as benchmarks for testing text classification models, such as BERT (Devlin et al., 2018) and XLNet (Yang et al., 2019).

Multimodal deep learning (Ngiam et al., 2011) is able to leverage different types of features, such as voice, image, and text, to achieve better performance. Nowadays, multimodal methods have been used in lots of tasks, for example, multimodal sentiment analysis (Soleymani et al., 2017), multimodal translation (Sanabria et al., 2018), multimodal emotion recognition (Tzirakis et al., 2017), and multimodal question answering (Yagcioglu et al., 2018).

One common multimodal architecture is to use different types of models to process the corresponding input data, such as first using an image classifier to obtain image features, a text classifier to obtain text features, and then combining these features before subsequent processing. In multimodal deep learning, the most critical part is feature fusion. Recent researches have proposed various feature fusion methods (Zhang

*Equal Contribution

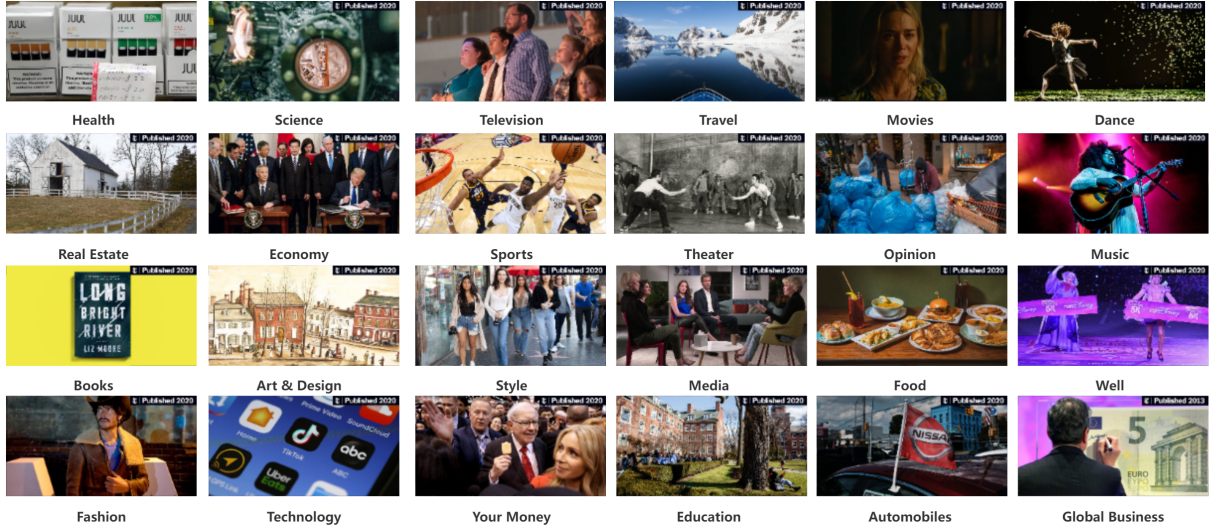


Figure 1: Image examples of 24 categories.

Category	Count	Category	Count
Health	3000	Books	3000
Science	3000	Art & Design	3000
Television	3000	Style	2681
Travel	3000	Media	3000
Movies	3000	Food	3000
Dance	3000	Well	681
Real Estate	3000	Fashion	3000
Economy	1761	Technology	3000
Sports	3000	Your Money	1263
Theater	3000	Education	825
Opinion	3000	Automobiles	1825
Music	3000	Global Business	1182

Table 1: Statistics of 24 categories.

et al., 2020a). *Concatenation* (Nojavanasghari et al., 2016; Anastasopoulos et al., 2019) is the most commonly used method. It splices different features directly along a certain dimension. Further, other fusion methods, for example, *weighted-sum* and *pooling*, are also able to achieve good results. *Weighted-sum* (Vielzeuf et al., 2018) assigns different weights to each feature and sum them up. *Pooling* (Chao et al., 2015) methods, including *max-pooling* and *average-pooling*, are also used in many fusion scenarios, which can find the most important pieces of information in each feature and finally integrate them. Additionally, *attention-based* fusion methods (Zhang et al., 2020b; Shih et al., 2016), which using the attention mechanism to let the model learn to automatically find the most crucial part of the feature through training, are playing an increasingly important role in multimodal deep learning tasks. Multimodal methods are also commonly used in news studies. Previous multimodal news researches mainly focus on fake news detection. Nakamura et al. (2019)

propose a multimodal fake news dataset from Reddit with six categories according to the degree and type of fake news in the news. Giachanou et al. (2020) use *word2vec* to extract the news text features and five different image models to extract news image features. Wang et al. (2018) use an adversarial neural network to identify fake news on newly emerged events in online social platforms. Fake news detection is a variant of news classification, which mostly has binary categories (true or false), making the task is not so difficult. Furthermore, there are few studies on the application of multimodal classification focus on real news.

In that case, we collect and apply multimodal methods on our dataset *N24News*, which containing massive news images and texts, as well as many different categories, to facilitate the research of multimodal news classification applied in real news study. The code and dataset will be on Github ¹.

3. The N24News Dataset

3.1. Dataset Collection

The *N24News* is extracted from the New York Times. New York Times is an American daily newspaper that was founded in 1851. It publishes worldwide news on various topics every day. Starting from the 2000s, the New York Times fully turned to digitization (Pérez-Peña, 2008), and previous news was transferred to the Internet to facilitate people’s reading and provide internet API for scientific research purposes.

To build the *N24News* dataset, we use the API provided by New York Times to obtain all the links published from 2010 to 2020. Then we use these links to retrieve all the actual web pages in the past decade. After analyzing those web pages, we exclude video news, and

¹<https://github.com/billywzh717/N24News>

Category: Movies

Headline: A Man’s Death, a Career’s Birth

Image:



Caption: Ryan Coogler on the BART platform at Fruitvale, where Oscar Grant III was killed. His film of that story, “Fruitvale Station,” opens next month.

Abstract: A killing at a Bay Area rapid-transit station has inspired Ryan Coogler’s feature-film debut, a movie already honored at the Sundance and Cannes film festivals.

Body: OAKLAND — It had been nearly a year since Ryan Coogler last stood on the arrival platform on the upper-level of the Fruitvale Bay Area Rapid Transit Station, where 22-year-old Oscar Grant III, unarmed and physically restrained, was shot in the back by a BART transit officer...

Table 2: A sample from *N24News*.

only the news articles in text form are retained. While most news has only one image, to better balance the number of images and news, we only choose one image for each news and drop out the news which does not contain any images. All news belongs to 24 different categories. We do not merge similar categories, such as science and technology, arts and theater. To make the dataset more balance, we collect up to 3000 samples for each category. Finally, 60K news articles are collected in total. The amount of each category is shown in Table 1. Each article sample contains one category tag, one headline, one abstract, one article body, one image, and one corresponding image caption. An example is shown in Table 2. We randomly split datasets into training/validation/testing sets in the ratio of 8:1:1. Compared with other multimodal research such as fake news detection, our dataset comes from a professional news website, which ensures the correctness of the dataset, thus no additional manual annotation work is needed.

3.2. Dataset Statistics

In Table 3, we show some information about *N24News* and other news-related datasets in previous researches. Compare to other datasets, *N24News* has some unique advantages. Firstly, *N24News* has 24 categories, which exceeds most of the previous news datasets, especially compare with multimodal datasets. Moreover, there is no valid multimodal news dataset that can be used to do real news classification before *N24News*. Previous multimodal researches in news classification mainly fo-

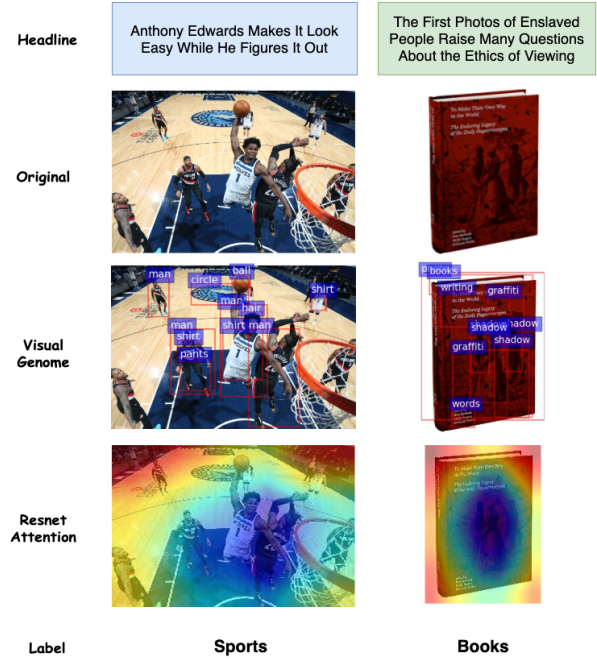


Figure 2: Visualization of critical parts in images. The news headlines lacks keywords that can be used for classification, but there are in the images.

cus on fake news detection with limited categories. The lengths of *Headline*, *Caption*, *Abstract* and *Body* are 52.33, 115.27, 129.42 and 4701.08 respectively. From *Headline* to *Body*, average lengths are progressively increasing. This allows us to study the gain effect of images on different lengths and different types of text classification tasks with *N24News*.

3.3. Multimodal Analysis

Text feature in classification task has been well studied before, so we will focus on what the news images in *N24News* are able to provide to improve the classification results. In Figure 1, we list some image examples of each category. It is obvious that news images are usually closely related to the category they belong to.

To better understand what can be learned from news images by current image classification models, we use a Faster-RCNN (Ren et al., 2015) model trained on Visual Genome (Krishna et al., 2016), a dataset aiming at providing semantic information from images. We also use a Resnet(He et al., 2016) trained on *N24News* to reveal the critical part of news images. Two examples are shown in Figure 2. Faster-RCNN extracts the important semantic information in the images, such as ball and books, and Resnet focuses on salient objects: player and book cover. In Figure 2, it is hard to recognize the topic only given the two headlines. The one on the left-hand side may be related to many topics, while the right-hand side one is closer to the topic of *opinion*. However, with the information obtained by images, we can easily guess that the left one is about a

Dataset	Size	Classes	Type	Source	Topic
20NEWS (Lang, 1995)	20,000	20	text	Newsgroup	real news
AG NEWS (Zhang et al., 2015)	1,000,000	4	text	AG News	real news
Guardian News (Hayat, 2018)	52,900	4	text	Guardian News	real news
Yahoo News (Yang et al., 2019)	160,515	31	text	Yahoo	real news
BBC News (Kaggle, 2018)	2,225	5	text	BBC	real news
BreakingNews (Ramisa Ayats, 2017)	110,000	none	text, image	RSS Feeds	real news
TREC Washington Post (Alexander et al., 2018)	728,626	none	text, image	Washington Post	real news
Fauxtography (Zlatkova et al., 2019)	1,233	2	text,image	Snopes, Reuters	fake news
Image-verification-corpus (Boididou et al., 2018)	17,806	2	text,image	Twitter	fake news
Fakeddit (Nakamura et al., 2019)	1,063,106	2,3,6	text,image	Reddit	fake news
N24News (Ours)	61,218	24	text, image	New York Times	real news

Table 3: Comparison of various news datasets.

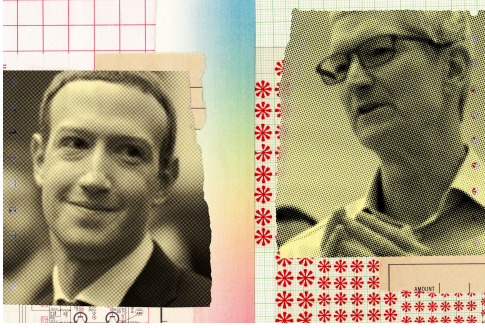


Figure 3: *Breaking Point: How Mark Zuckerberg and Tim Cook Became Foes*

sports-related topic, while the right one is about a book. Figure 2 shows that the information in the image can be used to strengthen news classification.

Moreover, The information provided by images can also help to distinguish similar categories. There are limited similar categories in previous news datasets. However, in *N24News*, there are some similar categories, for example, theater and movies. Only with text, it is difficult to tell the story is happening in a theater or on a screen, but images make things much easier. Theater-related images always happened on a stage, but movies not, as shown in Figure 1. This makes a huge difference, and if we can make good use of image information, the classification accuracy will be much higher.

3.4. Challenges

While multimodal data can introduce lots of new information to facilitate the news classification, *N24News* also releases some new challenges. The biggest challenge is how to better understand news images. Current image classification models are able to extract the features of objects and the relationships between them in images. However, directly using those models to

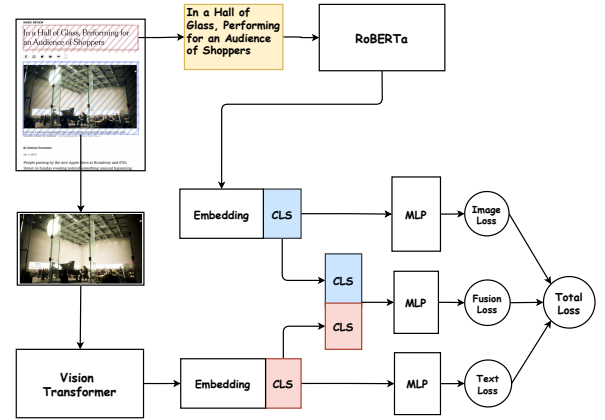


Figure 4: Overview of our multitask multimodal network.

classify news images cannot achieve a strong result because they are mainly designed to classify specific objects, such as cats or dogs, while a news image is more likely to reflect an event. An identical object may have different meanings in different scenarios. For example, the two people in Figure 3 are Facebook and Apple’s CEO, thus this image comes from a news related to technology. However, existing image models only recognize there are two people but cannot obtain more meaningful information. Therefore, the features obtained through those models cannot fully reflect the hidden contextual information in news images. We hope *N24News* can also prompt the research in event image classification, which is a new and challenging field.

4. Model

To figure out how images can enhance the news classification and the potential challenges when applying the multimodal methods, we use a simple multitask

Modal Type		F1	ACC
Image Only			
Image	-	52.80	54.34
Text Only			
-	Headline	70.31	71.98
-	Caption	71.56	73.87
-	Abstract	78.19	79.65
-	Body	87.65	88.86
Multimodal			
Image	Headline	78.42	79.41
Image	Caption	76.33	77.45
Image	Abstract	82.52	83.33
Image	Body	90.44	91.08

Table 4: The evaluation results on the *N24News* testing set.

multimodal network and conduct some experiments on *N24News*. As illustrated in Figure 4, our model consists of two kinds of feature extraction models. On the bottom is Vision Transformer (ViT) (Dosovitskiy et al., 2020), one of the current state-of-the-art image classification models. Above it is RoBERTa (Liu et al., 2019), one of the current state-of-the-art text classification models. The ViT we use is pre-trained on imagenet2012 (Krizhevsky et al., 2012), consists of a Resnet-50 and a base version of vision transformer with 12 layers transformer encoder. The pre-trained RoBERTa is also a base version and consists 12 layers transformer encoder.

We firstly use ViT and RoBERTa to obtain the image feature and text feature separately, where *embedding* is the embeddings extracted from the original text and image, *CLS* is a 1D embedding containing the information of its corresponding image or text. Then we concatenate those two kinds of features together. After obtaining the fused feature, we then use three multi-layer perceptrons (MLPs) to predict the label for image feature, text feature, fusion feature separately. Finally, the cross-entropy is used to calculate the *Loss* for each prediction. The final *TotalLoss* will be the sum of all three types of *Loss*. When testing on the test set, we only use the output of the fusion feature to calculate the final prediction result.

5. Experiments

5.1. Experimental Settings

We trained all the models in the *N24News* training set and the accuracy is tested on the testing set. Batch size is set to 32 and the learning rate is $1e-5$ with an Adam (Kingma and Ba, 2014) optimizer. Each input image is resized to 224×224 and the maximum length of each input text is set to 512. Training device is an NVIDIA Tesla V100 with 16 GB RAM. For each training process, we train the model with the training set, retain the model that performs best on the validation set, and apply it on the testing set.

5.2. Results

All the experiment results are shown in Table 4. We firstly classify the images and texts using ViT and RoBERTa respectively. In image classification task the accuracy is only 52.80 in F1, while RoBERTa behaves much better at the news text classification task. The lowest F1 is 70.31 using *Headline* and the highest F1 is 87.65 with *Body*. There is a direct correlation between RoBERTa classification accuracy and text length. From *Headline*, *Caption*, *Abstract* to *Body*, the longer the text length, the higher classification accuracy can be achieved. This is because RoBERTa can better understand the text with more meaningful words. It is found that the multimodal classifier is better than either the image classifier or the text classifier. Even for the *Body*, the improvement reaches 2.79 in F1 (**87.65** vs. **90.44**). This is powerful proof that multimodal learning combining image features and text features benefits news classification. And the result also shows that the shorter the text (from body to headline), the more obvious the gain effect of adding image features. In other words, when text contains insufficient information, image is a perfect supplement.

5.3. Error Analysis

To explore why the multimodal method surpasses the text-only method, we separate the trained baseline model into three types: original multimodal network, image classification network with only the ViT, and text classification network with only the RoBERTa. We then test them in the testing dataset using image-headline pairs. The experimental results are shown in Table 5.

It is evident that when image and text are both correctly classified, the multimodal network can nearly always classify news correctly. The correct-to-incorrect ratio is **42.46:0.03**. Additionally, when image and text are both wrongly classified, the multimodal network also tends to be incorrect, but the correct-to-incorrect ratio is **14.22:2.56**, much lower than the previous *Three True* situation. This shows that multimodal network can learn something useful after the features fusion of image and text, which may not be discovered if we process image and text separately.

Things are much more complex when only one of the image and the text classifiers is correct. The correct-to-incorrect ratio of multimodal classifier is $(27.69+7.01=\mathbf{34.7}):(2.40+3.63=\mathbf{6.03})$ in this situation. This shows that after proper training, the multimodal network will be more affected by the sub-network which can correctly perform the classification task. And this explains why our multimodal method is useful and able to outperform image-only and text-only networks.

The experiment results can be better understood by the examples in Table 5. It can be observed that images and texts can provide some complementary information. The multimodal method can thus classify news

Predict Result			Percent	Example		
Multi	Text	Image		Image	Headline	Label
True	True	True	42.46%		For Alphabet, a Record Fine Is Both a Footnote and a Warning	Technology
True	False	False	2.56%		Joan of Arc, Superstar? Not to the Woman Playing Her	Theater
True	False	True	7.01%		Time to Shift Gears	Food
True	True	False	27.69%		10 Dance Performances to See in NYC This Weekend	Dance
False	True	True	0.03%		His Dating Profile Listed Reasons Not to Date Him. She Was Intrigued.	Fashion
False	False	True	3.63%		Reading to Your Toddler? Print Books Are Better Than Digital Ones	Well
False	True	False	2.40%		Italy's Oldest Instrument Hints at Sounds of Prehistoric Rome	Music
False	False	False	14.22%		Tired of Mother's Day Brunch? Try a Tea	Travel

Table 5: The experiment results with three types of network. *True* means the classification is correct and *False* means the classification is incorrect. For each type of prediction, an example of corresponding image-headline pair and its ground truth label is provided. And the *Percent* represents each result ratio in testing set.

more accurately. In the third row, the topic of news may be easily considered about *Automobiles* if only considered the keyword *Shift Gears* in text. But when considering the image, the scene described in this image obviously talks about the food, not the car. On the contrary, in the fourth row, a group of people are performing on the stage. It is hard to categorize whether this news article belongs to *Dance* or *Theater* without texts. Luckily, the headline directly tells us they are dancing, and this article must belong to *Dance*.

Based on the above analysis, there are two main methods to further improve the performance of multimodal classification networks. The first one is to improve the behavior of each sub-network. If the accuracy (error) of sub-models is higher (lower), the multimodal prediction will also be improved. Our experimental results show that current state-of-the-art image classification models still have a long way to classify all news images correctly. The second method is to let the multimodal classifier be able to determine which sub-classifier extracts the more valuable feature. To do this, a more effective fusion network is needed to better combine image and text features.

6. Conclusion

In this paper, we introduce a multimodal news dataset *N24News*, which is collected from the New York Times containing both images and texts, enables the multimodal research in real news classification. Compared to previous datasets, it covers almost all the essential news categories in our daily life, making the research on it more applicable to the real world. Based on *N24News*, we propose a multitask multimodal network, which leverages the current state-of-the-art image classification model and text classification model. Experimental results show that combining image features and text features can achieve better classification accuracy comparing to the previous text-only methods. Our error analysis explains multimodal approach is helpful because the information of images and texts can complement each other. Accordingly, future work on improving the multimodal classification accuracy could include two main aspects: 1) improving image and text classification accuracy separately, especially the news image classification; 2) designing a more effective fusion network to better combine image and text features.

7. Bibliographical References

- Anastasopoulos, A., Kumar, S., and Liao, H. (2019). Neural language modeling with visual features. *arXiv preprint arXiv:1903.02930*.
- Chao, L., Tao, J., Yang, M., Li, Y., and Wen, Z. (2015). Long short term memory recurrent neural network based multimodal dimensional emotion recognition. In *Proceedings of the 5th international workshop on audio/visual emotion challenge*, pages 65–72.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Giachanou, A., Zhang, G., and Rosso, P. (2020). Multimodal fake news detection with textual, visual and semantic information. In *International Conference on Text, Speech, and Dialogue*, pages 30–38. Springer.
- 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.
- Katari, R. and Myneni, M. B. (2020). A survey on news classification techniques. In *2020 International Conference on Computer Science, Engineering and Applications (ICCSEA)*, pages 1–5. IEEE.
- Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M., and Fei-Fei, L. (2016). Visual genome: Connecting language and vision using crowdsourced dense image annotations.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. (2011). Multimodal deep learning. In *ICML*.
- Nojavanasghari, B., Gopinath, D., Koushik, J., Baltrušaitis, T., and Morency, L.-P. (2016). Deep multimodal fusion for persuasiveness prediction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 284–288.
- Pérez-Peña, R. (2008). Times plans to combine sections of the paper. *The New York Times. New*.
- Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 201.
- Shih, K. J., Singh, S., and Hoiem, D. (2016). Where to look: Focus regions for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4613–4621.
- Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S.-F., and Pantic, M. (2017). A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14.
- Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., and Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1301–1309.
- Vielzeuf, V., Lechervy, A., Pateux, S., and Jurie, F. (2018). Centralnet: a multilayer approach for multimodal fusion. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0.
- Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., and Gao, J. (2018). Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, pages 849–857.
- Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.
- Zhang, C., Yang, Z., He, X., and Deng, L. (2020a). Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, 14(3):478–493.
- Zhang, W., Tang, S., Su, J., Xiao, J., and Zhuang, Y. (2020b). Tell and guess: cooperative learning for natural image caption generation with hierarchical refined attention. *Multimedia Tools and Applications*, pages 1–16.

8. Language Resource References

- Alexander, Bondarenko, M., Völske, A., Panchenko, C., Biemann, B., Stein, M., and Hagen. (2018). Webis at trec 2018: Common core track. <https://github.com/irgroup/datasets/blob/master/WAPost/README.md>.
- Boididou, C., Papadopoulos, S., Zampoglou, M., Apostolidis, L., Papadopoulou, O., and Kompatsiaris, Y. (2018). Detection and visualization of misleading content on twitter. *International Journal of Multimedia Information Retrieval*, 7(1):71–86.
- Hayat, S. (2018). Guardian news dataset. <https://>

- //www.kaggle.com/sameedhayat/guardian-news-dataset.
- Kaggle. (2018). Bbc news dataset. <https://www.kaggle.com/c/learn-ai-bbc>.
- Lang, Ken. (1995). *Newsweeder: Learning to filter netnews*. Elsevier.
- Nakamura, Kai and Levy, Sharon and Wang, William Yang. (2019). *r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection*.
- Ramisa Ayats, A. (2017). Multimodal news article analysis. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 5136–5140.
- Sanabria, Ramon and Caglayan, Ozan and Palaskar, Shruti and Elliott, Desmond and Barrault, Loïc and Specia, Lucia and Metze, Florian. (2018). *How2: a large-scale dataset for multimodal language understanding*.
- Yagcioglu, Semih and Erdem, Aykut and Erdem, Erkut and Ikizler-Cinbis, Nazli. (2018). *Recipeqa: A challenge dataset for multimodal comprehension of cooking recipes*.
- Yang, Z., Xu, C., Wu, W., and Li, Z. (2019). Read, attend and comment: A deep architecture for automatic news comment generation. 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 5076–5088, Hong Kong, China, November. Association for Computational Linguistics.
- Zhang, Xiang and Zhao, Junbo and LeCun, Yann. (2015). *Character-level convolutional networks for text classification*.
- Zlatkova, D., Nakov, P., and Koychev, I. (2019). Fact-checking meets fauxtography: Verifying claims about images. *arXiv preprint arXiv:1908.11722*.