Analysing Narrative Manipulation in Russian Telegram News via Image and Text Divergences*

Maksym Bondar[1,2,*], Milita Songailaitė[1,2]

[1]Faculty of Informatics, Vytautas Magnus University, Universiteto 10 53361 Akademija, Lithuania
[2]Center for Applied Research and Development, Universiteto 10 53361 Akademija, Lithuania

Abstract

This paper aims to analyze narrative manipulation in Russian Telegram news channels by investigating the divergence between text and images. Through a detailed examination of selected Telegram channels, it was explored how images and texts are strategically manipulated to influence public perception and shape narratives. The study focuses on detecting inconsistencies in the visual and textual content in order to identify specific techniques used for disinformation. By analyzing patterns in image-text relationships, this research contributes the role of multimedia in forming public opinion and the manipulation of information in digital spaces.

Keywords

disinformation, narrative manipulation, russian telegram channels, image-text divergence, visual manipulation, textual manipulation

1. Introduction

Social media has become a powerful tool for spreading misinformation and manipulating narratives, particularly in the context of news dissemination [1, 2]. Studies highlight how platforms facilitate the rapid spread of fake news, disinformation, and misinformation, shaping public perception and influencing opinions [3]. In the case of Russian Telegram channels, this phenomenon plays a crucial role in narrative manipulation, as biased or deceptive content is strategically used to shape discourse and influence audiences [2]. This paper analyzes news from these channels, where texts are translated into English, processed, and compared with images using the BLIP model [4] and the similarity model SBERT [5], to explore the relationship between text and visuals in shaping narratives.

Understanding the mechanisms of visual communication—how visual conventions construct meaning and influence audience perception—plays a crucial role in resisting the manipulative power of visual media [6]. Scholars argue that visual literacy equips individuals with the ability to critically analyze images, making them more resilient against deceptive narratives propagated through television and digital platforms [7, 8]. In the context of Russian Telegram channels, where visual content is often strategically designed to shape public opinion, the ability to deconstruct these images becomes essential in identifying and countering narrative manipulation [9]. The role of framing in disinformation campaigns is also critical, particularly in how narratives are shaped and manipulated [10]. This perspective frames visual literacy as a defense mechanism against media-driven misinformation and will serve as the foundation for the following discussion.

2. Related Work

Deep learning and machine learning techniques provide powerful tools for analyzing narrative manipulation by detecting patterns between text and images. These methods help uncover how information is presented to influence audiences.

*IVUS 2025: Information Society and University Studies 2025, May 15, Kaunas, Lithuania
*Corresponding author.
maksym.bondar@vdu.lt (M. Bondar); milita.songailaite@vdu.lt (M. Songailaitė)
0009-0005-0239-284X (M. Bondar); 0000-0003-4315-2461 (M. Songailaitė)
© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-06-12

CEUR-WS.org/Vol-4213/paper24.pdf

Advancements in deep learning, driven by increased data availability and computational power, have led to significant progress across multiple domains. Deep neural networks have improved speech recognition [11], computer vision tasks such as image classification, segmentation, and action recognition [12, 13, 14], and natural language processing (NLP) applications like question answering [15], machine translation, and text summarization [16]. Many of these techniques can be adapted to analyze disinformation and narrative manipulation.

BLIP, a state-of-the-art vision-language pretraining model, performs well in image-text retrieval, captioning, and visual reasoning, making it highly relevant for media analysis [17]. According to Tan and Koehn [18], SBERT-based methods outperform LASER-based approaches, achieving approximately +2 BLEU improvement. This highlights SBERT's effectiveness in identifying and filtering high-quality sentence pairs.

Topic modeling is widely used to analyze text corpora, with methods including LDA [19], pLSI [20], NMF [21], and probabilistic models like CorEx [22]. Despite LDA's popularity, it struggles with short-text datasets, such as social media and microblogs [23, 24, 25, 26]. Alternative models, including DMM [27], DTM [28, 29], and BTM [30], have been introduced to improve performance on short and noisy texts [31]. However, most studies still rely on LDA, with a critical challenge being the manual selection of the optimal number of topics [32].

By integrating these methodologies, researchers can enhance their ability to detect manipulative narratives, offering deeper insights into media framing and disinformation strategies.

3. Data

3.1. Dataset of Messages

The dataset for this study was collected by crawling eight Russian Telegram[1] channels, extracting messages along with their metadata such as channel name, post date, message content, and message IDs. Each channel was processed separately to ensure structured data collection. It contains 18k rows of messages and images. The gathered dataset was then preprocessed by removing punctuation, converting text to lowercase, and eliminating stop words to enhance text analysis efficiency. Following preprocessing, the messages were translated into English using the *Helsinki-NLP/opus-mt-ru-en model*[2], preserving the original meaning while allowing for broader linguistic analysis. The final dataset retained both the original and translated texts, enabling comparative analysis of language structures and content across different narratives. Figure 1 shows the count of messages with their images per Telegram channel, with RVoenkor having the most messages in the dataset.

3.2. Dataset of Captions from Images

Captions for images were generated using the BLIP model and categorized by Telegram channel. By merging message IDs, the datasets were integrated to create a comprehensive multimodal collection linking text and visuals.

The LDA topic modeling results reveal main narratives within Russian propaganda Telegram channels, including themes such as military reporting, sentiment manipulation, technological warfare, geopolitical analysis, and election mobilization. Figure 2 illustrates the distribution of posts across these topics. Overall, the model identified five core themes consistently present in the analyzed channels.

The largest topic, Military Reporting, stands out as the most prominent theme. This topic emphasizes the portrayal of military operations, weaponry, and conflict-related content, which are strategically used to influence public perception and sentiment.

Table 1 presents the structure of the final dataset used in the study, detailing its components and categories.

[1]Accessible via https://telegram.org/

[2]Accessible via https://huggingface.co/Helsinki-NLP/opus-mt-ru-en

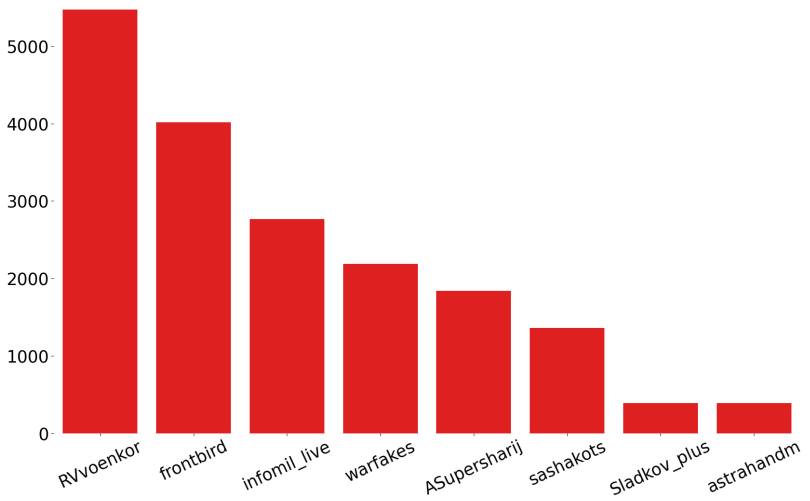

Figure 1: Count of messages per Telegram channel

Table 1
Final dataset

Column Name	Description
Author	Author of the post
Date	Date and time of the post
Message	Content of the message
Image Path	Path to the image associated with the post
Message ID	Unique ID of the message
Message Text	Processed text of the message
Translated Message	Translated version of the message
Image Name	Name of the image file
Description	Description of the image
Semantic Similarity	Semantic similarity score between message and image

3.3. Benchmark Dataset

The NewsMediaBias-Plus[3] dataset is a comprehensive collection designed for studying media bias and disinformation by combining both textual and visual data. It pairs news articles with corresponding images and includes annotations indicating perceived biases and content reliability. The dataset supports multimodal research on bias detection in media outlets. Annotations in this dataset are first generated by human annotators who label a subset of the data, followed by expansion via large language models (LLMs). These annotations are then validated and refined through a quality control process involving both automated checks and human reviews to ensure the accuracy of the labels.

For this study, the `image_description` column was selected as it provides textual descriptions of images paired with news articles. The goal of this approach is to evaluate the effectiveness of the narrative manipulation method by comparing these descriptions with captions generated by the BLIP model, using semantic similarity measures from SBERT. This allows us to assess how well the model captures the connection between the visual content and the textual descriptions provided in the dataset.

[3]Accessible via https://huggingface.co/datasets/vector-institute/newsmediabias-plus.

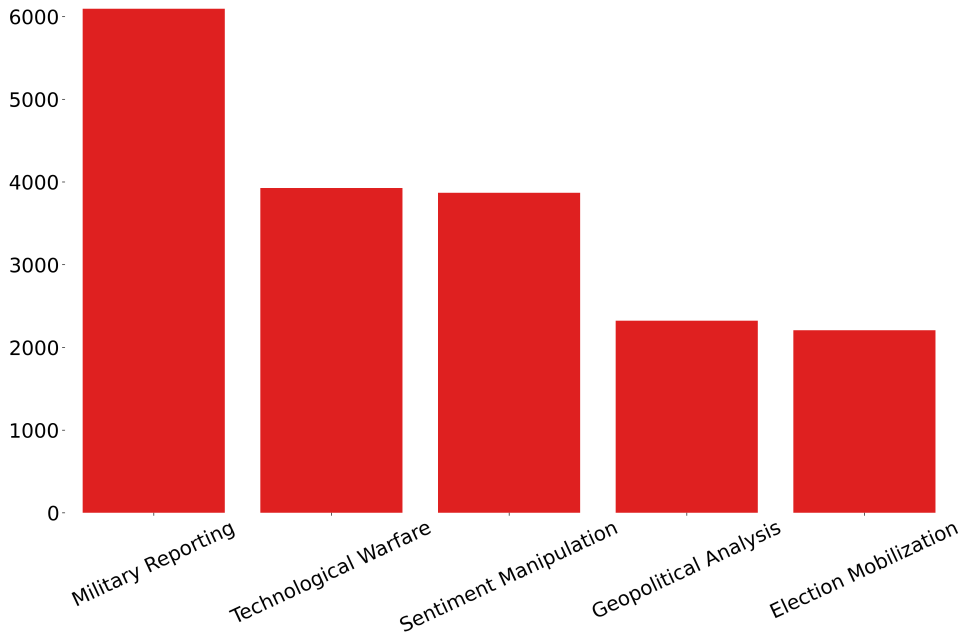

Figure 2: Visualization of the identified topics in the dataset using Latent Dirichlet Allocation (LDA)

4. Methods

4.1. BLIP Model

BLIP (Bootstrapped Language-Image Pretraining) is a vision-language model designed for multimodal tasks such as image captioning and retrieval. It consists of a Vision Transformer (ViT) for image processing and a Transformer-based text encoder-decoder (e.g., BERT/GPT) for textual understanding.

Given an input image I and an optional text prompt T, BLIP generates a caption or ranks candidate texts based on similarity. The Vision Transformer extracts visual embeddings:

$$V = f_{\text{ViT}}(I), \tag{1}$$

where $V = \{v_1, v_2, \ldots, v_n\}$ are visual feature embeddings.
If text $T = \{t_1, t_2, \ldots, t_m\}$ is provided, it is processed by a Transformer-based text encoder:

$$T' = f_{\text{BERT}}(T), \tag{2}$$

where T' represents the encoded textual embeddings.
The similarity between I and T is computed using cosine similarity:

$$S(I, T) = \frac{V \cdot T'}{\|V\| \|T'\|}, \tag{3}$$

where $S(I, T)$ measures alignment between vision and text embeddings.
BLIP is trained using three objectives:

1. **Contrastive Loss:** Maximizes similarity between correct image-text pairs while minimizing incorrect pairs:

$$\mathcal{L}_{\text{contrastive}} = -\log \frac{\exp(S(I, T^+))}{\sum_j \exp(S(I, T_j))}, \tag{4}$$

   where T^+ is the correct caption and T_j are all candidates in the batch.
2. **Matching Loss:** Ensures relevant text-image alignment.

3. **Captioning Loss:** Optimizes the likelihood of correctly generating text T:

$$\mathcal{L}_{\text{LM}} = -\sum_{i} \log P(t_i | V, t_{<i}), \tag{5}$$

where $P(t_i | V, t_{<i})$ is the probability of predicting the next token given previous tokens and visual features.

The total loss function is:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{contrastive}} + \lambda_2 \mathcal{L}_{\text{matching}} + \lambda_3 \mathcal{L}_{\text{LM}}, \tag{6}$$

where $\lambda_1, \lambda_2, \lambda_3$ are hyperparameters controlling the contribution of each loss term.

4.2. SBERT Model

SBERT (Sentence-BERT) is a modification of BERT designed for sentence similarity tasks. It fine-tunes BERT in a Siamese or triplet network setup to produce high-quality sentence embeddings.

Given two sentences S_1 and S_2, SBERT encodes them into fixed-length embeddings:

$$E_1 = f_{\text{BERT}}(S_1), \quad E_2 = f_{\text{BERT}}(S_2), \tag{7}$$

where E_1 and E_2 are the sentence embeddings.

Cosine similarity is used to compare sentence similarity:

$$S(S_1, S_2) = \frac{E_1 \cdot E_2}{\|E_1\| \|E_2\|}. \tag{8}$$

SBERT is trained with:

1. **Contrastive Loss:** Ensures similar sentences have closer embeddings:

$$\mathcal{L}_{\text{contrastive}} = (1 - Y) \max(0, m - S(S_1, S_2))^2 + Y S(S_1, S_2)^2, \tag{9}$$

where Y is 1 if the sentences are similar, 0 otherwise, and m is the margin.

2. **Triplet Loss:** Ensures a positive sentence S^+ is closer to an anchor S than a negative sentence S^-:

$$\mathcal{L}_{\text{triplet}} = \max(0, S(S, S^-) - S(S, S^+) + m). \tag{10}$$

3. **MSE Loss:** Minimizes the difference between predicted and actual similarity scores:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^{N} (S(S_1^i, S_2^i) - y_i)^2, \tag{11}$$

where y_i is the true similarity score.

The final loss function combines all objectives:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{contrastive}} + \lambda_2 \mathcal{L}_{\text{triplet}} + \lambda_3 \mathcal{L}_{\text{MSE}}. \tag{12}$$

To generate embeddings, SBERT applies mean or max pooling over token representations:

$$E = \text{Pooling}(\{h_1, h_2, \ldots, h_n\}), \tag{13}$$

where h_i are hidden states of the Transformer.

This allows SBERT to produce semantically meaningful sentence representations efficiently.

5. Results and Conclusions

5.1. Results

The semantic similarity analysis reveals a clear discrepancy between news post messages and their corresponding BLIP-generated captions, emphasizing the divergence between textual and visual content.

Figure 3 presents the results of this analysis, revealing that the overall similarity scores are low, which indicates a significant difference between the textual content and the corresponding images across most Telegram channels.

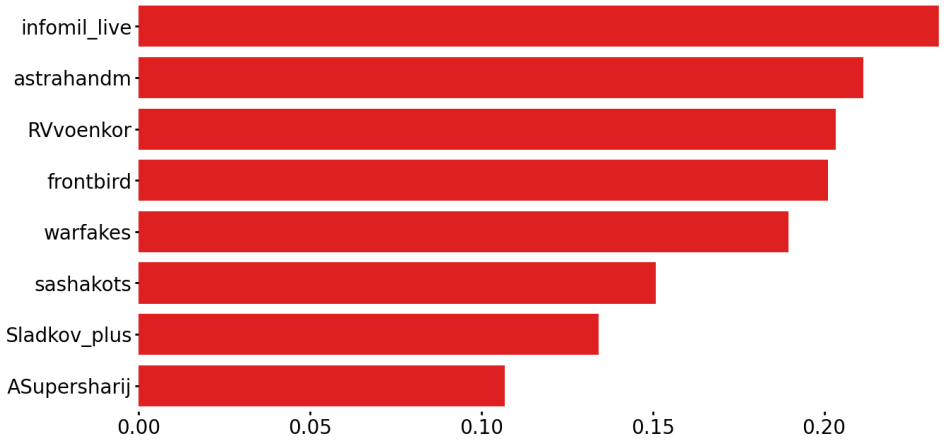

Figure 3: Semantic similarity between news messages and generated captions across the Telegram channels

Some channels, such as *infomil_live*, *astrahandm*, and *RVvoenkor*, have relatively higher similarity scores, while *ASupersharij* and *Sladkov_plus* show the lowest. These results suggest that the relationship between images and text varies across channels, with differences in how visual content complements or diverges from the accompanying narratives.

Table 2
Messages data with generated descriptions and semantic similarity

‘Message	Description	Similarity Score
"The missile track point was destroyed over the territory of the Belegorod region by the on-call means of the Ministry of Defence..."	a plane flying through the sky with clouds	0.097
"...Kiev region of explosions have previously been updated by two major explosions at the ground level in Kiev, and the enemy confirms no less than the arrival of the attacks..."	a rocket is launched by the russian space agency	0.28
"The enemy attacked a crossing plant in the Belgorod region, wounding civilians by droncomikadze,..., wounding three men..."	a red paint on the ground with a black line	0.198

Table 2 presents a set of messages from the *frontbird* and *RVvoenkor* Telegram channels, accompanied by descriptions generated by the BLIP model, and their corresponding semantic similarity scores. The BLIP model generates descriptions based on the context of the images associated with the messages.

The similarity scores are computed to measure how closely the generated descriptions align with the content of the messages.

Figure 4 illustrates a message reporting a missile strike in the Belgorod region by the Ministry of Defence. However, the accompanying image, showing a blue sky with some smoke, does not depict any missiles. This visual contrast with the described military action suggests a potential downplaying of the event's severity. The semantic similarity score of 0.097 further emphasizes the weak connection between the message and the image, highlighting the discrepancy between the text's narrative and the image's representation.

Next, figure 5 presents a message detailing explosions in the Kiev region, describing major ground-level explosions and enemy attacks. The accompanying image shows a rocket launched into the sky. While both the text and image involve military action, the image does not directly align with the described ground-level explosions. The semantic similarity score of 0.28 highlights this discrepancy, raising questions about the alignment between the narrative and the visual representation.

Lastly, figure 6 presents a message reporting an attack on a crossing plant in the Belgorod region, where civilians were injured by a drone strike. The corresponding image, showing red paint on the ground with a black line, does not directly relate to the described event. The image lacks any clear depiction of the attack or the resulting casualties, creating a noticeable contrast with the message's content. The semantic similarity score of 0.198 underscores the presence of mismatch between the text and image.

Figure 4: Sky with clouds

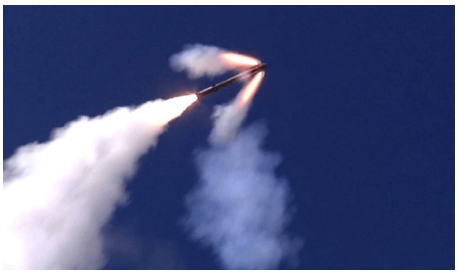

Figure 5: Launching a rocket

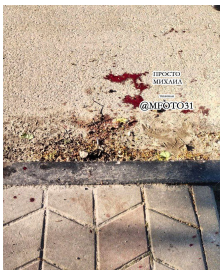

Figure 6: Blood on the road

To analyze narrative manipulation in Russian Telegram news, the semantic similarity of message-caption pairs in the custom dataset was compared to a benchmark of real annotated pairs (Figure. 7). High similarity scores were observed in the benchmark dataset, indicating strong alignment between text and images. In contrast, significantly lower and more varied similarity scores were found in the custom dataset, suggesting frequent mismatches.

Table 3
Model SBERT evaluation metrics

Metric	Value
Accuracy	0.9839
Precision	0.9770
Recall	1.0000
F1-Score	0.9884
Pearson Correlation	0.8153
ROC AUC	1.0000

In this study, the classification model showed strong performance in predicting similarity between the generated captions and the NewsMediaBias-Plus image descriptions. The classes were divided based on a cosine similarity threshold of 0.5: scores ≥ 0.5 were labeled as 1 (similar), and scores < 0.5 as 0 (dissimilar). The model achieved strong performance across various metrics, such as accuracy, precision, recall, and F1-score, as presented in Table 3. Furthermore, the Pearson correlation reflected a strong relationship between similarity scores and class labels, while the ROC AUC demonstrated perfect

discrimination. Figure 8 shows the confusion matrix with only a few misclassifications, confirming the model's high reliability and accuracy.

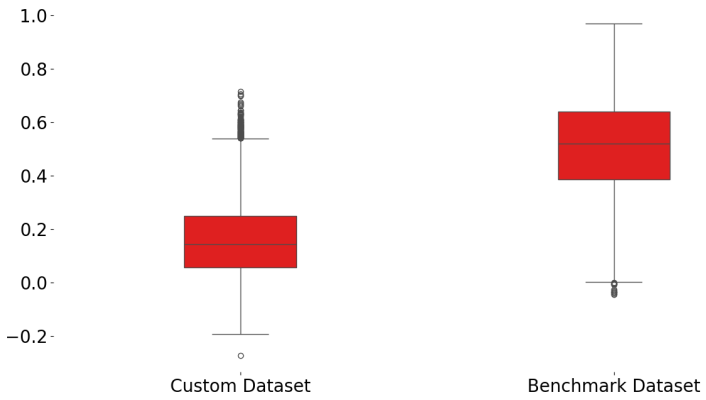

Figure 7: Semantic similarity comparison between the custom dataset and the benchmark dataset

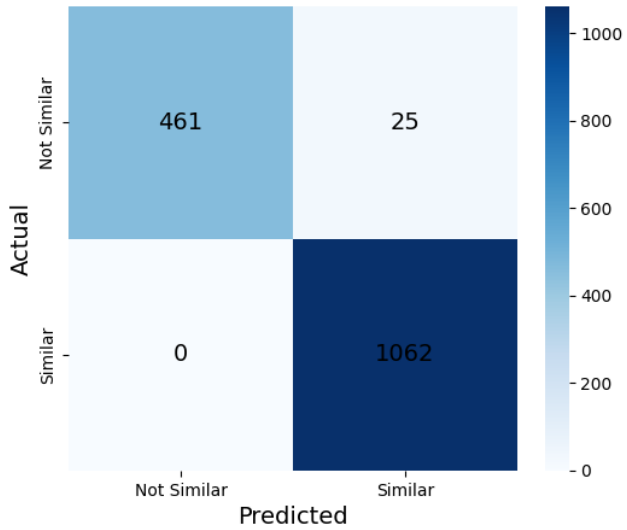

Figure 8: Confusion matrix of benchmark dataset

5.2. Conclusions

In this work, the study examines the role of image-text divergence in narrative manipulation within Russian Telegram news channels, utilizing BLIP for image-to-text captioning and SBERT for semantic similarity analysis.

1. Results show that visuals are often selected for emotional impact rather than factual accuracy, suggesting the use of propaganda, misinformation, or sensationalism.
2. Channels like *ASupersharij* and *Sladkov_plus* had the lowest semantic similarity scores, indicating higher divergence between images and text.
3. The benchmarking results show excellent model performance, with high accuracy, precision, recall, and F1-score. The strong Pearson correlation and perfect ROC AUC confirm the model's reliability, with minimal misclassifications.

Future work can focus on several key areas. Firstly, a comprehensive benchmark dataset will be created or identified to better evaluate models on image-text pairs. Additionally, more datasets will be collected from diverse news sources to expand the analysis of propaganda. To improve accuracy in detecting narrative manipulation, existing models will be enhanced, or more advanced models will be explored. The analysis will also be extended to multiple languages and regions to assess the global impact of narrative manipulation. Furthermore, the evolution of narrative manipulation over time will be investigated, particularly in response to geopolitical events.

Declaration on Generative AI

The authors employed Grammarly and QuillBot during the preparation of this work to ensure proper grammar and correct spelling. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] K. Sharma, et al., Fake news detection on social media: A data mining perspective, ACM Computing Surveys 51 (2019) 1–37.

[2] K. Shu, et al., Fake news detection: A survey, arXiv preprint arXiv:1708.01967 (2017).

[3] E. Aïmeur, S. Amri, G. Brassard, Fake news, disinformation and misinformation in social media: a review, Social Network Analysis and Mining 13 (2023). doi:10.1007/s13278-023-01028-5.

[4] J. Li, D. Li, C. Xiong, S. Hoi, Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, arXiv preprint arXiv:2201.12086 (2022). URL: https://doi.org/10.48550/arXiv.2201.12086, submitted on 28 Jan 2022 (v1), last revised 15 Feb 2022 (v2).

[5] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: https://arxiv.org/abs/1908.10084.

[6] J. A. Brown, Television "Critical Viewing Skills" Education: Major Media Literacy Projects in the United States and Selected Countries, Routledge, 1991.

[7] J. Rosenbaum, J. Beentjes, R. Konig, Mapping media literacy: Key concepts and future directions, Communication Yearbook 32 (2008) 312–353. doi:10.1080/23808985.2008.11679081.

[8] J. Lewis, The Ideological Octopus: An Exploration of Television and Its Audience, Routledge, 1992.

[9] D. Gomery, Media Ownership and Democracy in the Digital Information Age, University of Illinois Press, 1993.

[10] M. Haigh, Fighting and framing fake news, in: The SAGE Handbook of Propaganda, SAGE, 2020, pp. 303–323. Accessed: February 2025.

[11] A. Graves, A. rahman Mohamed, G. Hinton, Speech recognition with deep recurrent neural networks, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2013, pp. 6645–6649.

[12] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.

[13] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2961–2969.

[14] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803.

[15] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, Squad: 100,000+ questions for machine comprehension of text, arXiv preprint arXiv:1606.05250 (2016).

[16] R. Paulus, C. Xiong, R. Socher, A deep reinforced model for abstractive summarization, arXiv preprint arXiv:1705.04304 (2017).

[17] J. Li, D. Li, C. Xiong, S. Hoi, Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, arXiv preprint arXiv:2201.12110 (2022).

[18] W. Tan, P. Koehn, Bitext mining for low-resource languages via contrastive learning, Center for Language and Speech Processing, Computer Science Department, Johns Hopkins University (2021). URL: https://arxiv.org/abs/2109.05628.

[19] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, Journal of Machine Learning Research 3 (2003) 993–1022.

[20] M. Girolami, A. Kabán, On an equivalence between plsi and lda, in: SIGIR '03: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2003, pp. 433–434. URL: https://doi.org/10.1145/860435.860537. doi:10.1145/860435.860537.

[21] J. Wang, X.-L. Zhang, Deep nmf topic modeling, Neurocomputing 515 (2023) 157–173. URL: https://www.sciencedirect.com/science/article/pii/S0925231222012632. doi:https://doi.org/10.1016/j.neucom.2022.10.002.

[22] Q. Sun, Z. Yin, X. Li, Z. Wu, X. Qiu, L. Kong, Corex: Pushing the boundaries of complex reasoning through multi-model collaboration, arXiv preprint arXiv:2310.00280 (2025). URL: https://arxiv.org/abs/2310.00280. doi:10.48550/arXiv.2310.00280, cOLM 2024 / ICLR 2024 Workshop on LLM Agents.

[23] L. Hong, B. D. Davison, Empirical study of topic modeling in twitter, in: Proceedings of the First Workshop on Social Media Analytics, 2010, pp. 80–88.

[24] R. Mehrotra, S. Sanner, W. Buntine, L. Xie, Improving lda topic models for microblogs via tweet pooling and automatic labeling, in: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2013, pp. 889–892.

[25] X. Cheng, X. Yan, Y. Lan, J. Guo, Btm: Topic modeling over short texts, IEEE Trans Knowl Data Eng 26 (2014) 2928–2941.

[26] E. Jónsson, An evaluation of topic modelling techniques for twitter, 2016.

[27] J. Yin, J. Wang, A dirichlet multinomial mixture model-based approach for short text clustering, in: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2014, pp. 233–242.

[28] D. M. Blei, J. D. Lafferty, Dynamic topic models, in: Proceeding of the 23rd International Conference on Machine Learning, IEEE, 2006, pp. 113–120.

[29] E. del Gobbo, S. Fontanella, A. Sarra, L. Fontanella, Emerging topics in brexit debate on twitter around the deadlines, Soc Ind Res 156 (2021) 669–688.

[30] P. C. I. Pang, D. McKay, S. Chang, Q. Chen, X. Zhang, L. Cui, Privacy concerns of the australian my health record: Implications for other large-scale opt-out personal health records, Info Process Manag 57 (2020) 102364.

[31] J. Qiang, Z. Qian, Y. Li, Y. Yuan, X. Wu, Short text topic modeling techniques, applications, and performance: A survey, IEEE Transactions on Knowledge and Data Engineering (2020) 19.

[32] T. Griffiths, M. Steyvers, Finding scientific topics, Proceedings of the National Academy of Sciences of the United States of America 101 (2004) 9.