DOI:10.1145/3462671

The recent progress and ongoing problems of keeping online platforms and their users safe from malicious activities.

BY ALON HALEVY, CRISTIAN CANTON-FERRER, HAO MA, UMUT OZERTEM, PATRICK PANTEL, MARZIEH SAEIDI, FABRIZIO SILVESTRI, AND VES STOYANOV

Preserving Integrity in Online Social Networks

THE GOAL OF online social networks is to help create connections between people (online and offline), to connect people to communities of interest, and to provide a forum for advancing culture. Social networks advance these causes by providing a platform for free expression by anyone, whether they are well-known figures or your next-door neighbor. Unfortunately, open platforms for free expression can be used for malicious purposes. People and organizations can distribute misinformation and hate speech and can use the platform to commit crimes such as selling illegal drugs, coordinating sex trafficking, or child exploitation.

All these violations existed much before the advent of social networks, but social networks exacerbate the scale and sophistication with which these activities can be carried out.

Naturally, fighting these violations, which we collectively refer to as the problem of preserving integrity in online networks (or simply, integrity), has become a huge priority for the companies running them and for society at large. The challenges in preserving integrity fall into two general

» key insights

- **Designing policies for allowable content on social networks requires a balance between free expression and protecting users from harm.**
- **Detecting violating content with AI is an ongoing challenge because of the cultural, linguistic, subjective, and context subtleties.**
- **Effective techniques for enforcing integrity require advances in multimodal content analysis.**

IMAGE BY ANDRIJ BORYS ASSOCIATES. USING SHUTTERSTOCK

categories: policy and technical. Setting policies for what content and behavior are allowed on social networks is an area fraught with debate because it involves striking a balance between free expression and removing offending content. In addition, the policies must be sensitive to a variety of cultures and political climates all over the world. While we touch on the policy backdrop, this survey focuses on the technical challenges that arise in enforcing the policies. The technical challenges arise because deciding whether a post is violating can be extremely subtle and depends on deep understanding of the cultural context. To make things worse, content is created at unprecedented scale, in over 100 languages, and in very differing norms of social expression. In addition, preserving integrity is a problem with an adversarial nature—as the actors learn the techniques used to remove violating content, they find ways to bypass the safeguards.

The academic community has been actively researching integrity problems for the past few years and several surveys have been written about specific aspects of the general problem of integrity (for example, Pierri et al.[28] and Sharma et al.[33]). This survey comes from the perspective of having to combat a broad spectrum of integrity violations at Facebook/Meta.[a] The problems that Meta has had to tackle have also been experienced on other social networks to varying degrees.[b] The breadth of the services that Meta offers, the variety of the content it supports and the sheer size of its user base have likely attracted a widest set of integrity violations, and in many cases, the fiercest. This survey (and its associated longer version[11]) identifies a few sets of techniques that together form a general framework for addressing a broad spectrum of integrity violations and highlight the most useful techniques in each category.

a We refer to both Facebook and Meta in this article. Facebook refers to the application that is offered by the company, Meta.

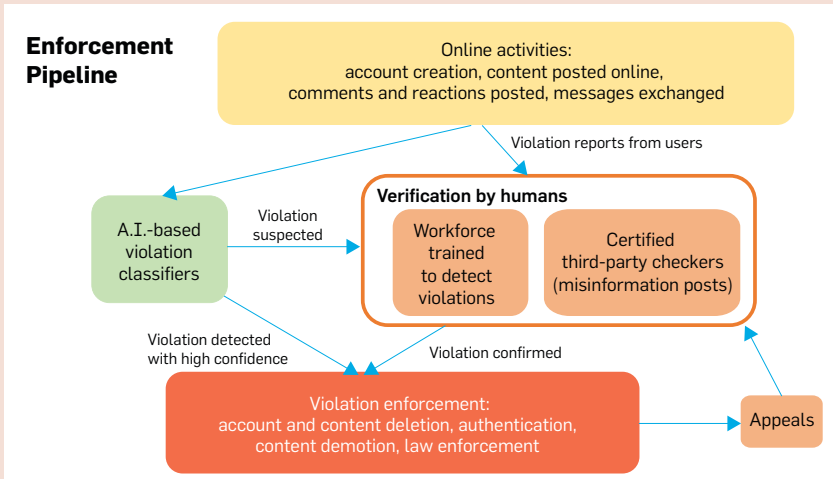
b While this survey is based on our experience at Meta, it is not meant to be a description of how Facebook tackles integrity or Meta's stance on integrity issues. When we use examples of policies or systems used at Meta, we call them out explicitly.

Problem Definition

This article considers mainly integrity violations that occur on social media posts, whether posted by individuals, groups, or paid advertisements. Messaging services that are offered by social networks have also been used as a vehicle for violating content policies, such as grooming children for future exploitation. As messaging services move toward end-to-end encryption, social networks need to find the right balance between privacy offered by the encryption and safety that can be further violated when messages are private. We touch on this issue briefly later.

Policies. The problem of preserv-

Figure 1. Enforcing integrity at Facebook.



ing integrity on social networks is defined by the community policies published by these networks that describe what is allowed on their platforms. While the final formulation of the policies is determined by the companies themselves, they are based on significant input from the community (for example, the European Commission) and local laws. The challenge in setting these policies is to balance free expression with the desire to keep the platform safe. For example, posting a bloody body is likely not allowed. However, if the context is a birth scene, then it could be allowed, if it's not showing private body parts. As another example, it is illegal to sell firearms or illegal drugs on Facebook, but users are allowed to debate the laws governing sales of these items. These subtleties make it even more difficult for an algorithm to decide whether a post is violating or not. It is also important to note these policies are not static. The policies and enforcement guidelines are updated as online discourse changes and new nefarious uses of social networks arise. For example, when a new type of misinformation surfaced that may result in physical harm (for example, bogus treatments against COVID-19), reviewer guidance was updated to be clear, a policy to remove content leading to imminent physical harm applied to this content.

Enforcing integrity. Figure 1 illustrates the flow of integrity enforcement at Facebook, which is similar in spirit to other social networks. Potential violations of integrity are detected in two main ways: reports from users who see the violating content and AI systems that inspect the content as it is uploaded. Content that the AI system deems violating with very high confidence may be immediately removed. When content is flagged, it may get demoted by the network to limit its virality while it is being verified. Demotion is done by downranking the content on users' stream so fewer people see it.

Potential violations are checked in two avenues depending on whether they are community standard violations or misinformation. In the former case, the content is sent to a large pool of paid content reviewers who are trained in the details of the violations prohibited by social network. Violating content will be removed if multiple reviewers agree that it is violating. In some cases, the reviewers are guided through a specific list of questions about the post that helps them make a justified recommendation.

Misinformation is treated differently because the social media companies do not feel they should decide what is true and false in the world.[c] Suspected misinformation violations are sent to third-party content reviewers. For example, in Meta's case, these reviewers are certified by an independent body—the International Fact-Checking Network. They review and rate content via primary and secondary research to find evidence that corroborates or does not corroborate a statement of fact. Because of the deeper nature of the research involved and because the number of third-party fact-checkers is much fewer, the throughput of the review pipeline is significantly smaller. In practice, many misinformation posts are small variations of previously debunked posts, and therefore significant effort has been devoted to finding semantically similar posts and to mapping new posts to previously debunked claims. When the machine learning model detects false content identical to content already rated by fact-checkers, it will apply fact-checks directly to the duplicate.

In addition to removing content, social networks can also reduce the distribution of content. In particular, the network can downrank content in users' news feeds or make it harder to reshare certain content, thereby reducing its distribution. Another method for protecting users is to make certain searches harder to conduct. For example, inappropriate interactions with children and selling of illegal goods often begin with bad actors searching for vulnerable individuals or users searching for products for sale.

Measurement. To assess the efficacy of the techniques for identifying violating content, we need a set of metrics we can track. Unlike many other machine learning applications, the adversarial nature of integrity and the fact that some violations happen with extremely low frequency, makes it tricky to design meaningful metrics. Meta publishes reports on its enforcement of integrity every three months based on several metrics.[8]

Prevalence is one of the key metrics of interest and is measured as a percentage of all content on the network. It refers to the amount of content on the social network and was not caught by the enforcement mechanisms. Prevalence, like recall of Web documents, can be tricky to measure, so it is typically done with respect to some sample. The simplest way to calculate prevalence is to count the

c Some misinformation, for example misinformation that contributes to the risk of imminent violence or physical harm, are covered by our Community Standards, and are not handled by this workflow.

number of distinct posts on the network. However, since some posts are viewed more than others, it is more meaningful to measure the prevalence of bad experiences, which is the number of times violating posts have been seen by users. Experience prevalence can also be refined to take into consideration the severity of the violation. For example, a completely nude photo of a person would be considered a more severe violation than a photo that has only a partial view of a naked body. Of course, considering severity requires there is a method to attach a severity measure to each post for each type of violation. In addition to the different forms of prevalence, other metrics that are tracked include: proactive rate, the percentage of violations that was detected by AI systems before users reported them; auto-deletion, the percent of posts that were deleted without human review; and appeals rate, that percent of posts that were deemed violating and the decision was appealed (and the appeal outcome).

Methods for Enforcing Integrity

We now describe methods for enforcing integrity. Rather than considering each violation type in turn, we identify the key aspects of the social-media ecosystem, each of which is common to a wide variety violation types. We focus particularly on analyzing the content of a post and on the interactions between users that ensued. Each of these topics represents an area for research and development, and the innovations that are found can be applied widely.

Text understanding. Semantic understanding of the text plays a key role in classifying whether a post is violating or not. Recent advances in self-supervised training[6,27,39] have shown great promise in addressing the thorny issues we encounter in social media posts, such as frequent misspellings, orthographic variations and colloquial expressions and the fact that the context of the words is important.

BERT[6] has become the standard architecture for accurate text understanding. Its recent refinements, such as RoBERTa,[19] ALBERT,[16] and T5,[30] have improved the training recipes and scaled to more data and parameters, thereby pushing the state of the art further. These approaches are especially helpful for difficult tasks like identifying hate speech because of the nuanced understanding of language that is required.

The multilingual challenge has been addressed as the problem of Cross-Lingual Understanding. In this setting, a model is trained to perform a task using data in one or more languages and is then asked to perform the task on data in other languages that are either not present or are underrepresented in the training data.

Self-supervised methods have been successful at tackling the multilingual challenge starting with multilingual BERT (mBERT).[6] mBert uses a single shared encoder to train a large amount of multilingual data. Further refinements, such as XLM[15] and XLM-R,[5] have closed the gap between in-language performance and performance on languages unseen during training data. XLM-R, in particular, has demonstrated that a multilingual model trained for 100 languages loses only a little in terms of accuracy (1.5% on average) when compared to a model specialized for a particular language on a variety of tasks. When applied to detecting integrity violations, these pretrained models are fine-tuned with training examples of individual violation types. Cross-lingual models such as XLM have been successfully used for problems like hate speech classification across languages.[1] In particular, some patterns that are used in hate speech posts do transfer across languages.

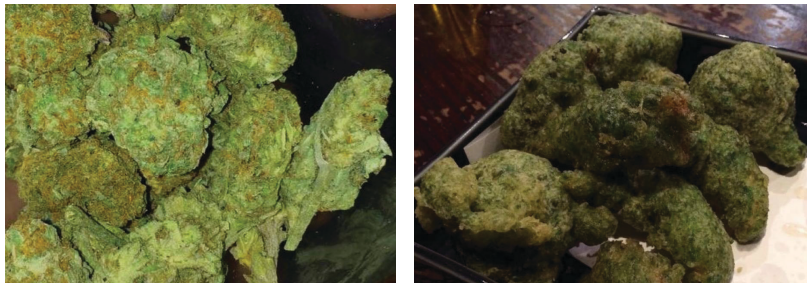
Nuances in text. A recent body of work considers whether how the content is conveyed can provide an important signal about the intent of the author.[31] For example, if the language used in a post involves emotions with high arousal (for example, anger), then that might indicate a more pronounced intention to hurt or mislead the reader. Other types of style analysis have also been used to detect fake news.[12,21]

Computer vision. Advances in computer vision have pushed the state of the art in supervised learning to a point where it is feasible to do image and video understanding with a high degree of accuracy. Figure 2 illustrates an example where current computer vision systems can distinguish between benign content (fried broccoli) and violating content (marijuana), where it would be arguably tricky for humans to do the same.

Analyzing video is a more challenging task since it requires understanding the semantics not only spatially, but also temporally and to consider audio when available. Often, videos are analyzed as 3D volumes (image+time) yielding compelling results.[35] However, the computational cost of analyzing 3D volumes has proven to be prohibitive and therefore researchers have developed techniques that factorize them into separate 2D convolutions (for images) and 1D convolutions (for time), while still achieving state-of-the-art accuracy.[35]

Manipulated media. Images, audio, and video shared across social media

Figure 2. Visual similarity between benign and violating content. One of the images is of marijuana and the other is of fried broccoli. The reader is encouraged to try to decide for themselves which of the images is the violating one. The correct answer can be found in the section "Emerging Topics and Challenges."



are rarely posted without some level of editing. With the recent progress in computer vision and speech processing and the advent of generative adversarial networks, manipulated media has leaped a significant step forward. These advances have unleashed potential for harmful applications, known as DeepFakes, such as face puppeteering,[23,40] speech manipulation,[29] face transfer,[34] and full body manipulation.[4] Many of these manipulations can be used to impersonate others, spread false information, or just introduce bias in the observer.

Accordingly, research on detecting media manipulation has become a prominent topic.[7,36] To date, two main techniques have emerged in this field. The first is based on learning the traces that media modification methods leave in the resulting footage[10,18] of the media. Different generation methods introduce subtle artifacts in the media derived from interpolation, inaccuracies of the generation process or the post-processing, usually hardly perceivable to the naked eye. Although these methods improve every year, such subtle artifacts are still one of main ways to attack detection of deepfakes. The second set of methods is based on analyzing physiological cues associated with human faces. Here, manipulated or full synthetic images lack some subtle cues like blood flow[25] or eye blinking,[17] which can be detected by computer vision methods.

Multimodal reasoning. Detecting many types of integrity violations, such as misinformation or hate speech, is often subtle because it is the combination of modalities that provides the real meaning of the content. As we illustrate with the memes in Figure 3, the text and the image in isolation can be benign, but when they are combined, their meaning changes and the content can become objectionable. For humans, understanding memes is easy, but for a machine it becomes harder than understanding each of the modalities alone. Unfortunately, in practice, many hateful posts are based on memes.

The current state of multimodal understanding is still in its infancy compared to our ability to understand each of the individual modalities. Hence, the field of integrity is an important impetus to pushing the state of the art on multimodal reasoning.

Recent research on the topic has started favoring classifiers based on early fusion over those based on late fusion. A late-fusion classifier uses existing unimodal classifiers and fuses them at the last layer. While they are simpler to build, they are ineffective at understanding content that combines multiple modalities in subtle ways. In contrast, early fusion classifiers feed the raw data into a fusion classifier before any predictions are made. One of the challenges in training classifiers that consider multiple modalities is that they are prone to overfitting to one of the modalities (for example, because that modality dominates the content).

Figure 3. In each of the three memes, both the text and the images, taken alone, are benign, but the combination results in an ill-intended meme.

Advances in representation learning also offer benefits to multimodal reasoning. The multimodal bitransformer[13] uses pretrained unimodal representations (BERT and ResNet-152) and then fine tunes them together for the task at hand. VilBERT (short for Vision-and-Language BERT),[20] on the other hand, pretrains the model using both text and image data, extending self-supervised methods so the system can learn early how text refers to parts of an image and vice versa. The multimodal bitransformer advances the more nuanced understanding at the category level, such as flagging entire classes of content about drugs or other harmful content. In contrast, VilBERT pushes the accuracy of multimodal understanding of object-specific tasks like question answering.

Based on these ideas, the Whole Post Integrity Embeddings System (WPIE) used in production at Facebook,[1] combines multiple sources of information in analyzing a post. WPIE is pre-trained across multiple modalities, multiple integrity violations, and over time. In a sense, just as cross-lingual pretraining can improve a classifier's overall performance, WPIE learns across dozens of violation types to develop a much deeper understanding of content.

Analyzing network behaviors. The analysis of network features plays a key role in recognizing violating content. In a sense, this should come as no surprise since the network is the medium used to disseminate and amplify the content and sometimes to modify the intent of the original post.

There are two aspects to analyzing network effects. The first is understanding user interactions with a post after it is published. Users interact with content through several mechanisms: reaction emojis (for example, like, ha-ha, angry, sad), commenting on a post, and resharing a post with their own network. As a result, content (violating or not) generates some reaction in the real world, and the nature of this reaction provides an important signal as to whether the content is violating. The second type of insights concerns understanding of the actors on the network (that is, users, groups, organizations) and relationships between them. For example, some actors may have been involved in

violations in other places and times on the network or may belong to communities prone to creating or even coordinating violating content.

Techniques for analyzing network behaviors are based on graph representations of the social network. The nodes in the graph correspond to users and to content, and the edges represent behavior in response to posts (for example, share, comments, reactions) and relationships between pairs of actors on the network. Some of the most effective works have considered automatically extracting user embeddings, by running graph learning algorithms such as Node2Vec[9] on those networks. Those embeddings are then used in downstream tasks such as that of detecting misinformation or abusive language detection[22,38] or hate speech. In a system that was used in practice in several integrity tasks at Facebook, Noorshams et al.[24] built a model that enriches the user embeddings with a temporal model of reactions to a post.

Coordinated action. Coordinated actions among multiple actors have become a common strategy to promote integrity violating content with possibly severe consequences. Coordinated behaviors refer to sets of actors that to either try to ensure a post gets wider distribution and appears more authoritative. For example, actors may band together to spread misinformation that election polling locations have closed earlier to prevent people from voting. Often, these coordinated actions are achieved through social bots.[2,32,37] Pacheco et al.[26] describe one method for exploiting the network structure to uncover coordinated activities within a social network, that is based on the surprising lack of independence of actions by users and exploiting this analysis to cluster together users that are likely to coordinate against some targets.

Emerging Topics and Challenges

Here, we briefly touch upon several emerging challenges.[d]

Integrity while maintaining privacy. In response to user sentiment about privacy, most messaging applications are moving to be encrypted from end to end. Consequently, since the content of the messages is no longer visible to the service provider, any analysis for integrity violations would need to be performed on the device itself. The machine learning model that is used for the inference would need to be trained offline on publicly available datasets and then shipped to devices. On-device inference is limited by the memory and processing power of the device as well as potentially having an adverse effect on its battery life. On-device inference also poses a challenge to measurement—the effectiveness of a model trained offline and shipped to clients will typically degrade over time, and if its performance cannot be measured it would be difficult to know when it needs to be fixed.

Advances in computer vision have pushed the state of the art in supervised learning to a point where it is feasible to do image and video understanding with a high degree of accuracy.

External knowledge can be extremely useful for detecting violations. As a simple example, in the context of misinformation, a system can try to decide whether a claim being made in a post is similar to one that has already been validated or debunked. Building on this idea, we believe that an interesting avenue for research is to endow integrity algorithms with broader knowledge about the real world and its subtleties. Indeed, deciding whether a post is violating often requires knowledge about current events, long-standing conflicts and troubled relationships in the world and sensitivities of certain populations.

Management of human content review workforce. One of the critical steps in enforcing integrity is inspection of content by human content reviewers who recognize violations. These reviewers receive content from either the enforcement systems or from users reporting violations, and the labels they produce are crucial to training the machine learning models that are used by the enforcement systems. In parallel, there is a set of content reviewers who work for third-party organizations and fact-check content suspected to be misinformation.

One challenge with operating such a workforce is the high volume of content that needs to pass their judgment. In addition, we need to be able to adapt and respond to scenarios when there is a sudden drop in workforce availability, as has happened during the COVID-19

d In Figure 2, the image on the left is fried broccoli and the image on the right is marijuana.

pandemic. The well-being of the content reviewers is a critical concern–the content can be graphic or otherwise objectionable. Operators of content moderation work forces may put several resources in place, such as providing access to licensed counselors, providing group therapy sessions, and screening applicants for suitability for the role as part of the recruiting process.

Conclusion

The prominence of social media as a medium for sharing news and information, as well as connecting people with friends and family, has grown in recent years. As a result, social media companies must balance between different goals: promoting engagement with the platform, creating meaningful interactions between users,[3] and protecting the integrity of online content. As this article has shown, from the technological perspective, preserving integrity presents a challenge that pushes on the boundaries of many aspects of artificial intelligence and its adjoining fields.

An important issue that integrity will soon grapple with is the boundary between content that obviously violates the policy and borderline content that may offend a wide audience. Social media platforms strive to minimize the number of negative experiences their users experience, and borderline content is a major source of such negative experiences. However, determining whether a piece of content will lead to a negative experience for a user is a highly subjective call and may be perceived as too much interference by the social media platform. Treading this fine line in a healthy fashion will surely be an important challenge in the upcoming years.

Admittedly, integrity is a tricky area for collaboration between industry and academia. Sharing datasets is problematic because of concerns regarding confidential user data, but also because some data is simply illegal to share (for example, a dataset of child exploitation imagery). Furthermore, some of the methods used to preserve integrity must be kept confidential, otherwise they can be weaponized by bad actors. Companies have invested significant effort to create datasets that can be used for research purposes, such as the Deepfake Detection Challenge[10] and the Hateful Memes Dataset.[14] We hope this article sparks ideas for more areas of possible collaboration. C

References

1. AI Advances to Better Detect Hate Speech, 2020; https://ai.facebook.com/blog/ai-advances-to-better-detect-hatespeech/.
2. Bovet, A. and Makse, H.A. Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications 10* (Jan. 2019).
3. Bringing people closer together, 2020; https://www.facebook.com/business/news/news-feed-fyi-bringing-people-closer-together.
4. Chan, C., Ginosar, S., Zhou, T., and Efros, A.A. Everybody dance now. In Proceedings of 2019 IEEE Intern. Conf. Computer Vision.
5. Conneau, A. et al. Unsupervised cross-lingual representation learning at scale. (2019); arXiv:1911.02116.
6. Devlin, J., Chang, M-W, Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. (2018); arXiv:1810.04805.
7. Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., and Ferrer, C.C. The DeepFake detection challenge dataset. (2020); ArXiv:abs/2006.07397.
8. Facebook Community Standards Enforcement Report, 2020; https://transparency.facebook.com/community-standardsenforcement.
9. Grover, A. and Leskovec, J. Node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD Intern. Conf. on Knowledge Discovery and Data Mining* (San Francisco, CA, USA, 2016). ACM, New York, NY, USA, 855–864; https://doi.org/10.1145/2939672.2939754
10. Güera, D. and Delp, E. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE Intern. Conf. Advanced Video and Signal Based Surveillance*. IEEE, 1–6.
11. Halevy, A.Y., Canton-Ferrer, C., Ma, H., Ozertem, U., Pantel, P., Saeidi, M., Silvestri, F., and Stoyanov, V. Preserving Integrity in Online Social Networks. (2020); https://arxiv.org/abs/2009.10311
12. Jeronimo, C., Marinho, L., Campelo, C., Veloso, A., and Sales da Costa Melo, A. Fake news classification based on subjective language. In *Proceedings of the 21st Intern. Conf. on Information Integration and Web-based Applications & Services*, 2019, 15–24.
13. Kiela, D., Bhooshan, S., Firooz, H., and Testuggine, D. Supervised multimodal bitransformers for classifying images and text. (2019); http://arxiv.org/abs/1909.02950
14. Kiela, D., et al. The hateful memes challenge: Detecting hate speech in multimodal memes. (2020); arXiv:2005.04790.
15. Lample, G. and Conneau, A. Cross-lingual language model pretraining. (2019); arXiv:1901.07291.
16. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. ALBERT: A lite BERT for self-supervised learning of language representations. In *Proceedings of the 2020 Intern. Conf. Learning Representations*.
17. Li, Y., Chang, M., and Lyu, S. In ictu oculi: Exposing AI created fake videos by detecting eye blinking. In *Proceedings of the 2018 IEEE Intern. Workshop on Information Forensics and Security*. IEEE, 1–7.
18. Li, Y. and Lyu, S. Exposing deepfake videos by detecting face warping artifacts. (2018); arXiv:1811.00656 (2018).
19. Liu, Y., et al. Roberta: A robustly optimized BERT pretraining approach. (2019); arXiv:1907.11692.
20. Lu, J., Batra, D., Parikh, D., and Lee, S. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of Advances in Neural Information Processing Systems 32: Annual Conf. Neural Information Processing Systems*. (Vancouver, BC, Canada, Dec. 8–14, 2019), 13–23; http://papers.nips.cc/paper/8297-vilbert-pretraining-task-agnostic-visiolinguistic-representations-for-visionand-language-tasks
21. Mihalcea, R. and Strapparava, C. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 Conf. Short Papers*. Assoc. Computational Linguistics, 309–312.
22. Mishra, P., Tredici, M., Yannakoudakis, H., and Shutova, E. Author profiling for abuse detection. In *Proceedings of the 27th Intern. Conf. Computational Linguistics*, 2018, 1088–1098.
23. Nirkin, Y., Keller, Y., and Hassner, T. FSGAN: Subject agnostic face swapping and reenactment. In *Proceedings of the 2019 IEEE Intern. Conf. Computer Vision*.
24. Noorshams, N., Verma, S., and Hofleitner, A. TIES: Temporal interaction embeddings for enhancing social media integrity at Facebook. (2020); arXiv:2002.07917.
25. Oh, T., et al. Learning-based video motion magnification. (2018); arXiv:1804.02684.
26. Pacheco, D., Hui, P., Torres-Lugo, C., Truong, B., Flammini, A., and Menczer, F. Uncovering coordinated networks on social media. (2020); arXiv:2001.05658.
27. Peters, M., et al. Deep contextualized word representations. (2018); http://arxiv.org/abs/1802.05365.
28. Pierri, F. and Ceri, S. False news on social media: A data-driven survey. (2019); http://arxiv.org/abs/1902.07539
29. Polyak, A., Wolf, L., and Taigman, Y. TTS skins: Speaker conversion via ASR. (2019); ArXiv abs/1904.08983.
30. Raffel, C., et al. Exploring the limits of transfer learning with a unified text-to-text transformer. (2019); arXiv:1910.10683.
31. Rajamanickam, S., Mishra, P., Yannakoudakis, H., and Shutova, E. Joint modelling of emotion and abusive language detection. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL, 2020.
32. Shao, C., Ciampaglia, G., Varol, O., Yang, K., Flammini, A., and Menczer, F. The spread of low-credibility content by social bots. *Nature Commun. 9* (2018).
33. Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., and Liu, Y. Combating fake news: A survey on identification and mitigation techniques. *ACM Trans. Intell. Syst. Technol. 10*, 3 (2019), 21:1–21:42; https://doi.org/10.1145/3305260
34. Thies, J., Zollhöfer, M., and Nießner, M. Deferred neural rendering: Image synthesis using neural textures. *ACM Trans. on Graphics* (2019).
35. Tran, D., Wang, H., Torresani, L., and Feiszli, M. Video classification with channel-separated convolutional networks. In *Proceedings of the The IEEE Intern. Conf. Computer Vision*.
36. Verdoliva, L. Media forensics and DeepFakes: An overview. (2020); ArXiv abs/2001.06564.
37. Vosoughi, S., Roy, D., and Aral, S. The spread of true and false news online. *Science 359* (2018).
38. Wu, L. and Liu, H. Tracing fake-news footprints: Characterizing social media messages by how they propagate. In *Proceedings of the 11th ACM Intern. Conf. Web Search and Data Mining*, 2018, 637–645.
39. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, 2019, 5754–5764.
40. Zeng, X., Pan, Y., Wang, M., Zhang, J., and Liu, Y. Realistic face reenactment via self-supervised disentangling of identity and pose. (2020); ArXiv abs/2003.12957.

Alon Halevy is Director at Meta AI, Menlo Park, CA, USA.

Cristian Canton-Ferrer is Research Manager at Meta AI, Seattle, WA, USA.

Hao Ma is Director at Meta AI, Seattle, WA, USA.

Umut Ozertem is Senior Staff Software Engineer at Google, San Francisco, CA, USA.

Patrick Pantel is Director at Meta AI, Seattle, WA, USA.

Marzieh Saeidi is a Research Scientist at Meta AI, London, U.K.

Fabrizio Silvestri is a professot at Sapienza University, Rome, Italy.

Ves Stoyanov is Applied Research Manager at Meta AI, Menlo Park, CA, USA.