No.46
January 2021

HardenStance Briefing

Trusted research, analysis & insight in IT & telecom security **PUBLIC/UN-SPONSORED**

A Way to Turn the Tide on Fake News

Imperva's CTO, Kunal Anand, has a compelling idea for amassing enough AI brain power and investment dollars in one place to help turn the tide in the battle against fake news.

- Fake news is undermining democracy and literally killing people. Finding better solutions for mitigating it must be a priority for the incoming Biden Administration.
- Joining forces to invest in classifying fake news is one way social media companies could drive efficiencies and faster response times in addressing this problem. The technology could be openly licensed out to support automated policy enforcement.
- Actions like new Antitrust suits against Facebook are transforming the landscape, making one or more new directions all but inevitable. Social media companies should be motivated to think out of the box to respond to this political pressure.

Fake News Will Be the Death of Many More of Us

Notwithstanding the fantastic, good use that many people make of it, there is a profound crisis in the impact social media is having around the world. This is exemplified by recent Pew Research Centre survey data conducted in July and then published in October 2020. This found that 64% of American adults surveyed believed social media "have a mostly negative effect on the way things are going in the country today. Only one in ten said social media sites "have a mostly positive effect on the way things are going."

HardenStance takes the view that fake news is a serious threat – ultimately even an existential threat – to human societies.

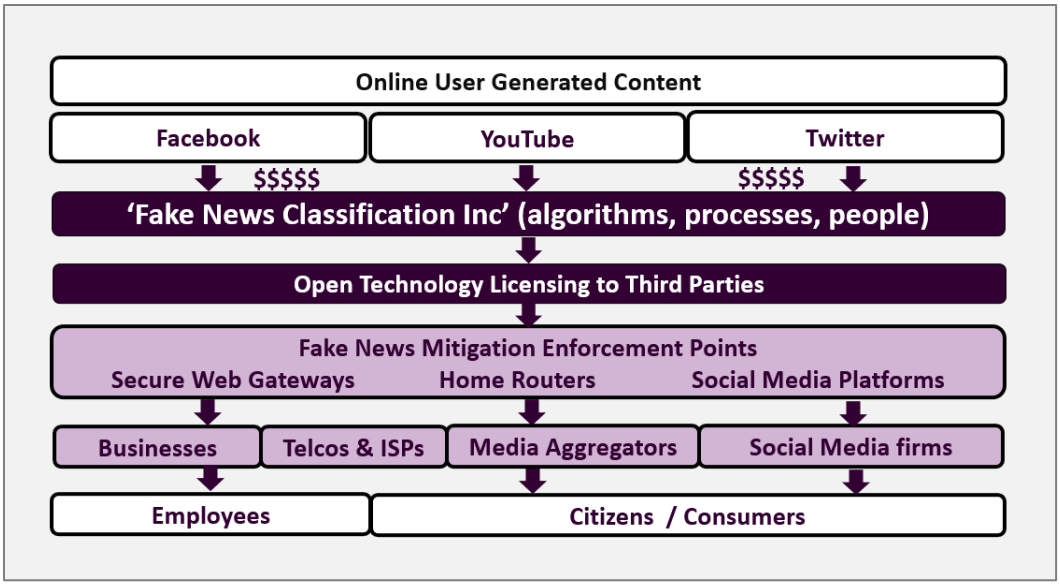
The rapid rise of fake news – information that is either false or misleading – is one of the primary factors driving this disaffection with social media. HardenStance takes the view that fake news is a serious threat – ultimately even an existential threat – to human societies. The most obvious examples include the highly effective efforts to discredit the facts and impact of climate change as well as the COVID-19 pandemic. U.S. President Donald Trump's use of fake news to try and manipulate the outcome of the recent Presidential election also shows its pivotal role in undermining democratic pluralism.

There are exhaustive debates to be had over the optimal rules for governing social media companies and how the fake news that proliferates on their platforms should be addressed. These debates include, but are not limited to, how fake news should be classified; how rules and regulations for mitigating fake news should be defined and who by; what those rules and regulations should be; and how these rules should be enforced. Just before Christmas, Federal and state prosecutors in the U.S. filed landmark antitrust suits against Facebook which threaten the nuclear option of breaking the company up. These are complex philosophical or political issues which this Briefing does not address.

A Fix that's Independent of Governing Philosophy

Drawing on a recent conversation with Imperva's CTO, Kunal Anand, this HardenStance Briefing focuses on the state of the art in technology and business models in the way fake news is classified. It looks at the way rules around blocking, disrupting, and mitigating the impact of fake news are generated and enforced. And it points to a candidate model for developing cutting edge solutions that could potentially improve the scope, speed, and efficacy of fake news mitigation. This model is purposely intended to be independent of the incentives driving any of the social media organizations as well as adaptable to any governing philosophy mandated by government regulation.

Figure 1: A Model for a Jointly-Funded Fake News Classification Start-Up



Source: HardenStance/Imperva

This Briefing does take one philosophical stance, albeit it's hardly a controversial one. It assumes that the big social media companies should take primary responsibility for identifying and classifying fake news. This is on the grounds that they are first to see new content. Identifying and classifying fake news is already part of what these companies do, albeit with limited success currently.

The social media companies themselves should take primary responsibility for identifying and classifying fake news.

Kunal Anand's idea is captured by HardenStance in **Figure 1**. He advocates that the big social media companies should jointly invest in standing up their own company – let's call it 'Fake News Classification Inc'. The idea is that this company should jointly develop the very best processes and advanced algorithms and pool teams of content moderators.

Such a company would be better resourced and better placed than any one organization to identify and classify fake news accurately and quickly enough for it to be acted upon almost as soon as it's posted. Kunal's idea is that this jointly-funded company should then license its technology out openly to other players in the ICT ecosystem. These can then serve as fake news mitigation enforcement points.

The model depicted in **Figure 1** takes no philosophical stance on whether the creation of such a company should arise from the enlightened self-interest of the social media giants, through indirect government pressure, or via direct government mandate. Neither does this Briefing make assumptions around which actors in the ecosystem should take responsibility for what kind of enforcement actions – or indeed whether such actions should be undertaken voluntarily or via a regulatory mandate.

Multiple Motivations Driving Different Players to Want to Act

This model merely recognizes that different actors could be motivated to act as enforcement points in the battle against fake news. For example, most businesses have their own ethical codes around people viewing sexually explicit, violent, racist, homophobic or sexually prejudicial content via any of their organization's IT assets.

If these organizations could have easy access to sufficiently accurate filtering capabilities, it's reasonable to suppose that some would want to extend those policies to block, label or otherwise distinguish fake news. Telcos and ISPs are potential enforcement points too, although again that could be via a government license obligation or as a self-interested exercise in market differentiation. Media aggregators could also leverage this technology to differentiate themselves. The social media companies should have their own motivations.

Rationale and Contours of a Jointly Funded Company

So what's the rationale for a hypothetical 'Fake News Classification Inc' funded by Facebook, YouTube, and Twitter - and what would such a company look like? Before becoming CTO of Imperva, Kunal Anand was Director of Security for Myspace from 2006–2009. Hence, he has some direct experience in this area and maintains strong contact with peers that are familiar with the state of the art in mitigating fake news.

At root, social media is the ultimate in trusted environments. There is no authentication of any kind between the producer and consumer of content. There is no baked-in mechanism for authenticating that a piece of content even originated from the source it purports to originate from, let alone validating the content itself.

Intermediaries like Apple News compound the problem by lending what are trusted brands to aggregating content, none of which is inherently trustworthy. When content like a web page is updated or revised, users are oblivious to those changes. Originators of posts that embedded the first version are left vulnerable to association with revised content that they no longer support or with an original post which is no longer valid. A viable long term fix to this problem requires a roadmap for evolving from a wholly trusted social media environment to something resembling the Zero Trust environment that the enterprise IT world is striving for.

While it may be celebrated as a great leap forward in the artificial creation of content with a language structure, GPT-3 will be just as useful for generating superb quality fake content.

The tools available for generating good and bad, real and fake, information, are also becoming increasingly sophisticated. For example, the AI world is bristling with excitement about the Open AI foundation's new GPT-3 algorithm. While it may be celebrated as a great leap forward in the artificial creation of content with a language structure, GPT-3 will be just as useful for generating superb quality fake content.

"It's the People, Stupid – not just the Stupid People"

The increasing speed, scale and accuracy with which automated Bots can both generate and disseminate fake news is, of course, a fundamental part of the problem. Facebook has already taken down literally billions of fake Bot accounts, for example.

But the way human beings respond to fake news is also a key part of the problem:

- A Massachusetts Institute of Technology (MIT) study of 2018 found that fake news travels roughly six times faster than accurate news. Critically, it found that while bad Bots disseminate accurate and inaccurate information at roughly the same rate, humans are much more likely to retweet false information than accurate information. This is because false information is more likely to trigger a reaction of disgust or anger which is more likely to inspire onward forwarding.
- It's a misnomer that better fake news controls are needed mainly to give 'stupid people' better tools to distinguish fact from fiction. Many 'smart' people mistakenly believe they will always know fake news when they see it. But spotting fake news is already difficult and the challenge will get even harder as Bots get more advanced. Recognising the limitations on our ability to spot fake news means accepting that our own personal information security 'perimeter' will inevitably be breached.
- Social media posts are increasingly considered news in themselves. The outgoing US President's use of Twitter has made the most obvious contribution to this phenomenon. Even HardenStance has become unwittingly dragged into this. HardenStance's Tweets are now routinely published as supporting analyst quotes in news reporting by a variety of news media.

State of the Art Detection and Mitigation is Highly Manual

The best way to think of the state of the art in fake news detection, classification and mitigation is to observe that the problem is substantially unsolved despite Facebook outsourcing the work of content moderation to no less than 15,000 individuals via third party contractors. To make a bigger dent in addressing the problem, a June 2020 report from New York University's Stern Centre for Business and Human Rights advocated doubling that number to 30,000 content moderators. The report also advocated bringing all these contractors in-house to include them as a core function of Facebook's business.

Machine learning algorithms have certainly improved in terms of their ability to identify factual inaccuracies. They've also improved in terms of their ability to identify key words as well as correlate key words and flag issues accordingly. Despite these improvements, the self-evident truth is that content moderation is still heavily dependent on human intervention for policy enforcement. Automated labelling and blocking of fake news is still in its infancy. Since it is able to proliferate at such a huge scale in seconds, human intervention necessarily allows (sometimes literally) fatal delay in policy enforcement. Clearly, what's needed is increasingly automated enforcement in real-time or near real-time of the kind that only sophisticated algorithms can enable.

Increased levels of intervention in recent weeks may look like significant progress. Notably Facebook and Twitter have started labelling President Trump's tweets with "this claim about election fraud is disputed". However, this too relies on manual intervention. These policies are only applied (by humans) to the tip of the iceberg of very senior, very high profile, individuals by the social media platforms. The vast majority of users remain largely free to spread misinformation. In fact, as widely reported in the closing weeks of 2020, Facebook has been formally operating a whitelist of 110,000 U.S. government officials and candidates that formally exempts their posts from any fact-checking.

If you want to track the history of a social media post under the hood to rate its authenticity almost instantaneously, it's going to take world class AI algorithms to do that.

Could the Stern Centre's proposal of 30,000 moderators make a bigger dent in Facebook's fake news universe? Probably. But would 30,000 really minimize the risk from fake news to the trivial level we need to get it down to? Can you really scale filtering to the right level of granularity and accuracy with more humans? Given that the tools available for AI-assisted, fake news creation are advancing so fast, it's clear that investment in AI-assisted fake news classification and automated enforcement per **Figure 1** has to play a much bigger role. Ultimately, if you want to automatically spot an AI-created post or track the history of a post under the hood to rate its authenticity almost instantaneously, it's going to take world class AI algorithms to do that.

A Collaborative Model

As stated, the social media companies are, of course, already investing in people and tools to combat fake news within their own siloed platform environments. But consider this: at $70.7 billion in 2019, Facebook's annual revenues were twenty times Twitter's at $3.5 billion. It's therefore likely that Facebook is already spending more than Twitter's entire annual revenue on fake news classification and policy enforcement. But on its own, Facebook isn't even close to solving the problem. Hence, even if Twitter were to channel every cent of its revenue into it, it couldn't solve this problem alone either.

This is where the logic of Kunal Anand's idea of investment by the social media companies in a jointly-owned 'Fake News Classification Inc' kicks in. Until now these firms have accepted that they have to do 'something' – and be seen to do 'something' – about fake news. And they have indeed done 'something'. But social and political pressure is mounting and telling them that 'something' is no longer enough. They need to be doing – and be seen to be doing - everything they can to put brilliant people to work to get a grip on this toxic problem at the same time as they grow their businesses.

The model Kunal Anand is proposing is a decentralized one by design. It allows any new classification technology that's jointly developed to be made openly available to any policy enforcement player that wants it. This would support the kind of flexibility needed

in a global Internet environment which is already highly fragmented, if not balkanized. It would also support a model of enforcing government-mandated policies around fake news that needs to be adaptable to new social, political and technology developments.

The landscape is quite clearly shifting all around the social media companies. Their leaders have a track record of being myopically focused on one another as competitors. They should recognize now that joining forces to arrive at a more automated and scalable solution to this problem is in each of their company's own self-interest. ■

More Information

- Contact HardenStance's Principal Analyst: patrick.donegan@hardenstance.com
- Register here for free notifications whenever HardenStance releases new content.
- www.hardenstance.com
- HardenStance received no payment – direct or "in kind" – for publishing this briefing.

About HardenStance

HardenStance provides trusted research, analysis and insight in IT and telecom security. HardenStance is a leader in custom cyber security research and leading publisher of cyber security reports. HardenStance is also a strong advocate of industry collaboration in cyber security. HardenStance openly supports the work of key industry associations, organizations and SDOs including NetSecOPEN, AMTSO, The Cyber Threat Alliance, The GSM Association, OASIS, and ETSI. www.hardenstance.com

HardenStance Disclaimer