Project Name: Working for Democracy

Reviewing Twitter Blue in South Africa

Release date: 09 June 2023

Introduction

In a Twitter post that was created in March 2023, Billionaire Elon Musk, who according to Forbes now owns 74% of Twitter[1] explained that the only realistic way to address advanced AI bot swarms taking over the social media site is to ensure that only verified accounts will be eligible to appear in the "For you" recommendations page - the first page that appears when you sign in to your Twitter account. This post came late at night on the same day that Musk had also tweeted around 4 a.m to explain that a paid verification system would increase the cost of bot accounts by approximately 10 000% i.e. making it impractical from a cost perspective for the creators of bots to continue operating on the platform.

A bot, broadly speaking, is an automated program used to engage on social media that is simple to manage with a single person being able to operate hundreds or even thousands of accounts[2]. On Twitter, the bot problem takes the form of accounts posing as people with strong political opinions and have been credited with issues like election interference[3]. Understanding that changes were going to be made to the Twitter platform on 15 April 2023 to minimise the risk of bots, researchers at the CABC noticed that the bot indicator score of an account that we found posting content of interest that we then shared in a report about the National Shutdown, @Nhleiks5, jumped from a 30% likelihood of the account being a bot to 70%. While there are multiple factors that contribute to the calculation of a bot indicator score - a score assigned to a Twitter account by sites like Truthnest and Botometer that allow journalists and analysts to rapidly assess the likelihood that an account may be a bot - this prompted us to review other accounts that have been historically problematic for South African Twitter based on reports that the CABC has put together since 2020 to understand if the changes that were made - as the platform switched to a subscription based model to authenticate accounts - had had an impact on scores like the bot indicator.

The switch to a subscription based model requires the input of credit card details for the activation of a Twitter Blue account. This is powered by the popular paygate called Stripe, which connects online platforms with easy to set up payment solutions. The implication here is that once the credit card details are known, the complete anonymity of the person or people behind the accounts should be lost as the details connected to that credit card could be used, to an extent, to track down the account holder that may be behind an account that posts

[1] https://www.forbes.com/profile/elon-musk/
[2] https://www.cloudflare.com/learning/bots/what-is-a-social-media-bot/
[3] https://www.alestlelive.com/opinion/article_b95367da-e96b-11eb-ae0c-cb6db8f80f05.html

inflammatory content or hate speech. Once your card details are in, the platform requests that you enter a registered mobile number in order to verify your account.

With the Promotion of Access to Information Act (PAIA) in place along with the implementation of POPIA (Protection of Private Information Act), the Regulation of Interception of Communications and Provision of Communication Related Information Act (RICA) and the Financial Intelligence Centre Act (FICA,) the CABC sought to understand if and how these details could be accessed to tighten the clamp on peddlers of distorted narratives and vitriolic content.

Accordingly, in the sections that follow we provide a contextual analysis of Twitter's switch to a subscription model for verification. First, we explore the changes that were made to Twitter and how it has impacted users. Second, we consider the implication of the changes based on the South African legal framework. Third, and finally, we consider the effectiveness of the Twitter changes by analysing accounts that contribute to the South African Twitter conversation on key issues of concern for citizens - i.e. accounts that now have blue ticks on their account - to consider if there is any notable difference in the style and content they are posting since joining Twitter Blue.

The Changes

In March 2023, a senior correspondent at online news site Vox claimed that Twitter had become a degraded product since Musk took over the company, citing major glitches and platform downtime as one of the signs of the degradation of the product[4]. In the interest of objectivity and fairness (which characterises our approach towards analysing and understanding the large data volumes that are processed by CABC), we noted that this was not the first time that Twitter had experienced long periods of downtime. A 2016 headline read "Twitter suffers longest outage in its history"[5]. In October 2020, another headline read "Twitter Down: Social Network Suffers Widespread Technical Problems"[6]. Then, in July 2022, "Twitter experiences longest global outage in years"[7].

The restoration of accounts that had been previously suspended for violating the terms on Twitter did not bode well for Musk. While neo-Nazi and QAnon accounts were reinstated, the accounts of journalists that critiqued Elon Musk were suspended[8]. These actions flew in the face of Musk's "free speech absolutism" as he allegedly

[4] https://www.vox.com/technology/2023/2/16/23603155/elon-musk-twitter-worse-degrading-quality-glitches-superbowl-boost-feed
[5] https://eandt.theiet.org/content/articles/2016/01/twitter-suffers-longest-outage-in-its-history/
[6] https://variety.com/2020/digital/news/twitter-down-widespread-outages-1234806792/
[7] https://www.theguardian.com/technology/2022/jul/14/twitter-experiences-longest-global-outage-in-years
[8] https://www.vox.com/technology/2023/2/16/23603155/elon-musk-twitter-worse-degrading-quality-glitches-superbowl-boost-feed

requested the alteration of posts, i.e. it was reported that at the time of the Super Bowl President Joe Biden made a post that received more engagement so, Musk had engineers at Twitter alter the post positioning so that his post was more prominent than that of Biden's. It has also been alleged that tweets and accounts that criticised the Indian government had been blocked at their behest[9].

What is Twitter Blue and how has this changed Blue tick verified accounts?

Twitter Blue is a subscription based service that allows users with a verified phone number and credit card details to sign up and receive the benefits that are shown in Figure 1. Not only does a blue tick verified account now mean that subscribers will see less paid for advertising on the platform; changes to the Twitter functionality will also now prioritise the ranking of a post by subscribed accounts so that they are featured more prominently than accounts that are not subscribed.

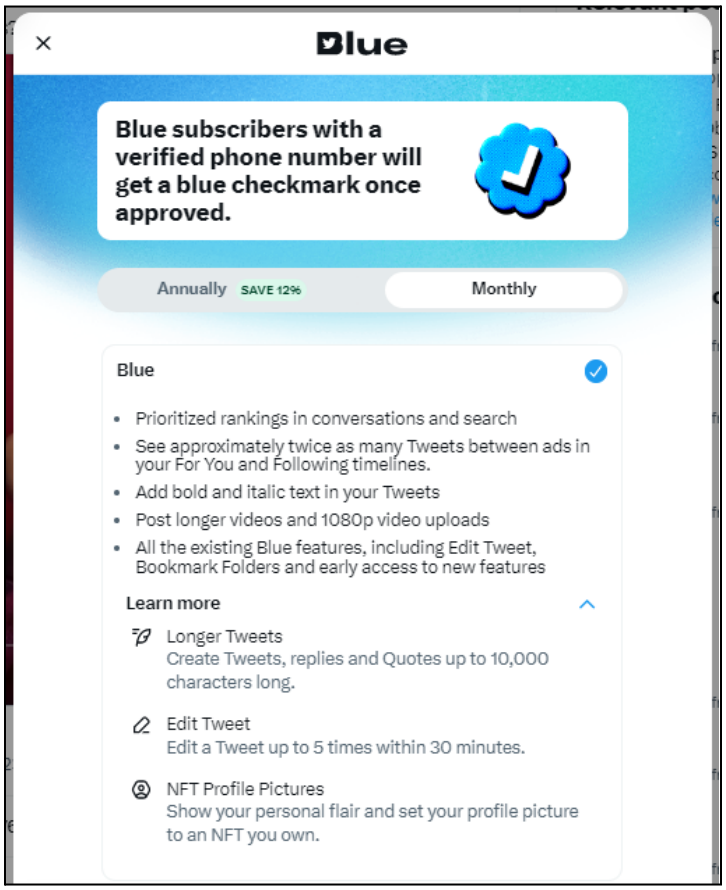

Figure 1: Twitter Blue value adds

Organisations paying a higher subscription fee and those that are verified receive a yellow/gold tick next to their Twitter name. The accounts of known government officials and official organisations, like the account of

[9] https://www.theatlantic.com/ideas/archive/2023/04/elon-musk-twitter-free-speech-matt-taibbi-substack/673698/

President Cyril Ramaphosa, received a grey tick next to them as the changes occurred. Examples of organisation and government ticks are shown in Figure 2 below.

Figure 2: Examples of yellow ticked organisations and grey ticked public official accounts

This can be seen as a big step in the right direction for some as the old Twitter Blue tick verified account system could only be given to celebrities, influencers, journalists and people who were meant to be active, notable and of public interest. As a result, currently, authenticity is granted to those with a registered mobile number and credit card details. With anti money-laundering and consumer protection acts like FICA, RICA and POPIA, one would expect that tracking peddlers of disinformation and hate speech should become easier for law enforcement authorities. However, the details of these Acts and some key limitations are explained below.

It may be of interest to note that while reviewing Twitter Blue changes on the recommendations “for you” page that content from accounts that were not subscribed to Twitter Blue still appeared in the feed. Using a newly created account, with no interactions and only a few people selected to follow, posts from accounts with no blue ticks, addressing politics and other issues of social concern, appeared on the home feed, meaning that their content has somehow been prioritised over other content. This calls into question the effectiveness and validity of the new changes, begging further investigation into how and why certain content is prioritised over others. Please see Figure 3 for an example of posts by accounts with no blue ticks that featured high on the news feed of the newly created account.

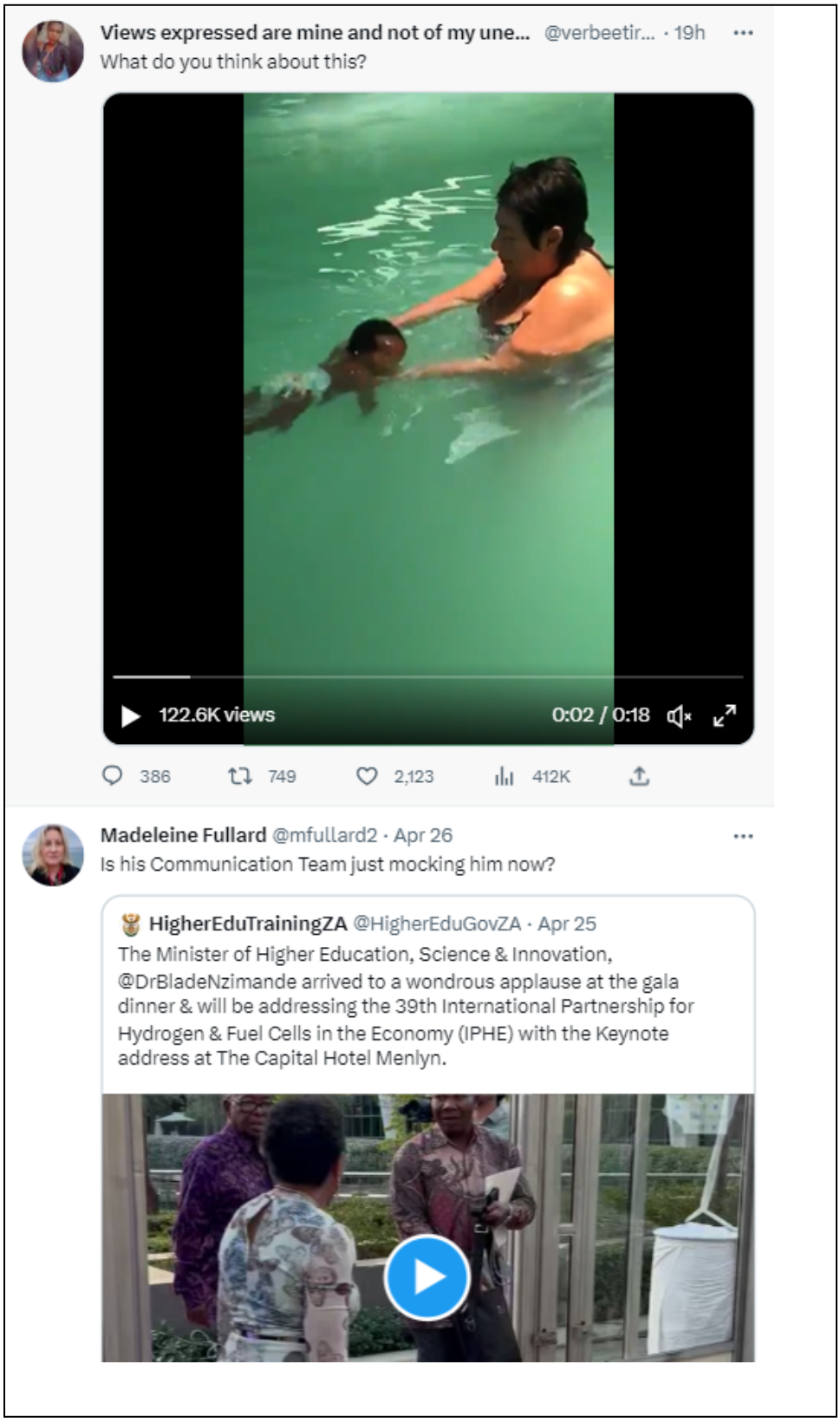

Figure 3: Content from non blue ticked accounts that were given prominence on the "for you" home feed.

The Law

Regulating online influencers has proven to be a particularly hard task across the globe.[10] Legislatures - particularly in Europe - have responded to public interest challenges accompanying the burgeoning technology and social media platforms. In South Africa, we have a host of legislation specifically dealing with data and the interception of data communications, such as the POPI Act, PAIA, ECTA and RICA. The first step in taking legal action lies in identifying the actor behind the account. Often, and in spite of our legislative framework, this remains a challenging task.

Ordinarily, the Constitution - along with legislation such as the POPI Act - protects individuals with privacy rights and places an obligation on platforms to refrain from processing or availing private information about an individual without their consent and/or cause. An added layer of protection may be added by the user themselves, who may create social media accounts with a throwaway email address or use VPNs (a Virtual Private Network that creates an alternate IP address that can be based anywhere in the world) to protect their real identity. In addition, certain social media platforms similarly place a high value on the privacy of their "clients" and add additional safeguards to protect their data. Naturally, all laws have their limitations and data protection and privacy are no different in this respect. Moreover, legislation is generally playing catch-up with the fast pace of the evolution of digital technologies and online social networking, and how these are co-evolving to yield impacts that are of public interest.

As noted above, the POPI Act is a law that protects people's personal information. This includes the use of personal information like your name or email address or other personal information when creating an account on social media sites like Twitter. However, there are some exceptions to the general protections. If a public organisation needs to use your personal information to prevent or detect illegal activities, they can sometimes bypass the protections in the Act. Social media corporations such as Twitter can also be given permission to use your personal information if it's in the public interest. For example, if a government agency suspects someone of breaking the law online, they may be allowed to access that person's personal information.[11] Similarly, an online site may avail your personal information if the public interest outweighs your right to privacy. The Information Regulator, an independent body established in terms of section 39 of the POPI Act, is empowered to monitor and enforce compliance with the Act and indeed, PAIA.

PAIA similarly deals instead with access to information. The Act was introduced to give effect to section 32(1)*(a)* & *(b*) of the Constitution, which provides everyone with the right to access any information held by the State or by another person where that information is required for the exercise or protection of any rights. [12]

[10] C Peter 2020 *The Regulation of Social Media Influencers.*
[11] The Act defines a public body as any department of state or administration in national, provincial or local spheres, as well as functionaries or institutions exercising a public power/function or performing a duty in terms of the Constitution.
[12] Act 2 of 2000, s 50.

Notably, where a public body requests access to a private body for the exercise or protection of any rights, it must be acting in the public interest - and not its own.

RICA regulates the interception of communication and the provision of communication-related information and may be of further relevance here. The Act has its limitations. It contains references to repealed legislation and terms, such as Internet Service Provider, that are defined differently in other legislation such as ECTA. It is also vague in parts which makes it a rather difficult piece of legislation to interpret and apply. Nonetheless, the Act requires electronic communications service providers to request and keep records of their customers when entering into service provision contracts.[13] Further, the Act governs the interception of data messages. This allows law enforcement to acquire any identifying information from private communications made by an account through the lawful procedures set out in the Act.[14] Certain sections of the Act were declared unconstitutional by the Constitutional Court in *Amabhungane Centre for Investigative Journalism v Minister of Justice and Correctional Services*, where the court questioned whether adequate safeguards exist to justify the extent of intrusion by the Act. The declaration of invalidity was suspended in three years in the 2021 judgment, to give Parliament adequate time to rectify the Bill. The court did not, however, declare the entire Act unconstitutional, only certain sections that failed to provide adequate protection of privacy and freedom of expression.

Having laid out the legislative framework[15] within which one can gain access to personal information of a social media account, one must turn to consider the efficacy of these laws. The first procedural hurdle is perhaps the most obvious one - legislation is *generally* territorial. This means that the legislation mentioned above operates within the borders of South Africa only. As such, natural persons within the borders are liable for their actions on social media but the entity itself, without a legal presence in South Africa, does not necessarily have to conform to the laws of the country despite the applicability of these laws to it. This is due to the fact that social media organisations such as Twitter lack a legal presence in South Africa, making it incredibly difficult to enforce the data protection framework on the corporation.

The second hurdle relates to Twitter's policy on availing personal information to interested parties. Twitter's guidelines stipulate that they will only reveal the identity of a user if they are required to do so "by appropriate legal process such as subpoena, court order, or other valid legal process documents".[16] South African legislation governing social media can, in fact, be said to apply to Twitter and other social media corporations to the extent that their activities affect South Africa. This requires recognition of a South African subpoena, court order or other valid legal process document. The relevant chosen legal process will have to be

[13] Act 40 of 2002, s 39 & 40.
[14] Ibid at section 2 through to 19.
[15] The Electronic Communications and Transactions Act 25 of 2002 will be dealt with below.
[16] Twitter Guidelines for Law Enforcement. Available at https://help.twitter.com/en/rules-and-policies/twitter-law-enforcement-support

recognized by a US court or alternatively, proceedings would need to be instituted in a US court through a US firm.[17] Once these requests are validly obtained and Twitter recognizes the legal request, the organisation can provide the influencer's name, email address, IP address, and other identifying information that could assist law enforcement authorities to clamp down on purveyors of disinformation.

As noted above, the recent change to avail the blue-tick to everyone at a cost has had its fair share of critics and indeed, proponents. It is interesting to note that the blue-tick is not "verification proper". For instance, celebrities and notable persons were given the tick to prevent catfishing or impersonations of their account to spread false information. Now, you can have a quasi-impersonating account and still have the blue tick. In essence, Patricia Lewis need not have her real name (or one commonly associated with her) on Twitter to obtain the blue tick. All that is required is payment of the monthly subscription fee to Twitter. This activity, in and of itself, may allow for the easier identification of these "verified" accounts.

When activating Twitter Blue, you are required to input your financial information, which includes your initials, surname, card number, expiry date and card type (i.e. VISA or MasterCard for example). This information is processed by popular online payment gateway, Stripe Inc. Procedural hurdles notwithstanding, Twitter Blue's effect is to allow actors to identify accounts much more quickly than previously. The Electronic Communications and Transactions Act[18] regulates the acquisition of personal information obtained through electronic transactions and provides that a registry must be created for the storing of personal information obtained through electronic transactions. This Act similarly provides that a data controller (in this case, Stripe as the custodian of the banking data) may disclose personal information obtained if required or permitted by law.[19] Interestingly, the aforementioned provisions are not automatically legally binding on a data controller.[20]

Twitter Blue does provide an additional layer of identification and verification but it is not a panacea for the challenges of identifying social media accounts. Its efficacy in regulating social media personalities depends on its use and how effective it is in deterring bad actors on the platform. In the following section, we look at whether Twitter Blue has had any effect on the behaviour of verified personalities.

The reality

[17] L Michalson *'Impersonation and fake social media accounts'* available at https://www.michalsons.com/blog/impersonation-fake-social-media-accounts/9732.
[18] Act 25 of 2002.
[19] Ibid at s 51(6).
[20] Section 50(2) of the ECTA provides that these principles must be voluntarily subscribed to by a data controller by recording such facts in an agreement between the two parties.

Twitter's new eligibility criteria[21] for receiving the blue checkmark states that, in addition to subscribing, accounts need to have: a display name; profile picture and confirmed phone number; and be active for the past 30 days. Accounts should also show no signs of being "misleading or deceptive".

The criteria state that, "You may not misappropriate the identity of individuals, groups, or organisations or use a fake identity to deceive others. We want Twitter to be a place where people can find authentic voices.

"While you are not required to display your real name or image on your profile, your account should not use false profile information to represent itself as a person or entity that is not affiliated with the account owner, such that it may mislead others who use Twitter," according to the terms and conditions.

In previous reports, the CABC has outlined a number of features that could lead to an account being deemed suspicious (i.e. indicating inauthentic behaviour). Among these are:

- Using a 'fake' image on their profile or an image that may belong to someone else;

- Having a high daily tweet volume (an average tweet volume being between 2 - 3 times per day based on a recent study); and

- A high retweet volume with little original content; among others.

On the one hand, the above features may indicate suspicious and even bot accounts. Alternatively, parody accounts could also use some of these features without necessarily being regarded as inauthentic. Twitter's request is for these accounts to specify that they are in fact parody, fan or commentary accounts.

Earlier this year, the popular @ChrisExcel102 account topped the trending lists after a woman claimed to be the face used by the account. In an Instagram post, Bianca Coster wrote: "Over the past few years I've been ridiculed, insulted and just downright tormented because of the Chris Excel account on Twitter. I chose to keep silent and never to address it formally in hopes that this frenzy would eventually die out. This silence has been misconstrued as a lot of things, most annoyingly being that I am in compliance with this bully".

Coster added that the damage caused by the account had been humiliating and emotionally draining. "Constantly having to validate my innocence and disassociation with the account. Before there was Chris

[21] https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts

Excel, there was me. A person in the flesh with dreams, goals and a full life to live. It's my turn to tell my story & in that hopefully bring awareness towards cyber-bullying".

There has been a mixed bag of online responses, from users who criticised the verified account for hiding behind a fake identity to those claiming to stand by the account, despite the backlash. In response, the account (which had 1.6 million followers at the time of this report) tweeted that Chris Excel "is a brand, and it's not affiliated with any person".

The post read: "Not even a single tweet that came forward saying Chris is a female. This is a brand that stands with black men and brothers. It's been reported many times, and it's still standing."

Two days later, @ChrisExcel102 shared a screenshot stating that the account that had been reported multiple times, did not violate Twitter policy - which states that:

- Parody, commentary or fan accounts that "depict another person, group, or organisation in their profile to discuss, satirise, or share information about that entity" are not in violation of the misleading and deceptive identity policy.

"While these accounts may use elements of another's identity, they also include profile language or other indicators that inform people on Twitter that the account is not affiliated with the subject of the profile."

It further adds that, in order to avoid confusing others about its affiliations, these accounts must distinguish themselves in their account name and in their bio.

"The account name should clearly indicate that the account is not affiliated with the subject portrayed in the profile. Accounts can indicate this by incorporating words such as, but not limited to, 'parody,' 'fake,' 'fan,' or 'commentary.'"

"...The bio should clearly state that the account is not affiliated with the subject portrayed in the profile. Non-affiliation can be indicated by incorporating words such as, but not limited to, 'not affiliated with,' 'parody,' 'fake,' 'fan,' or 'commentary.'"

Despite including "commentary" in its bio and allegedly not being in violation of Twitter policy, the @ChrisExcel102 handle has and continues to use the image of Bianca Coster on its profile and to post derogatory content online - especially about women and celebrities (Figure 4).

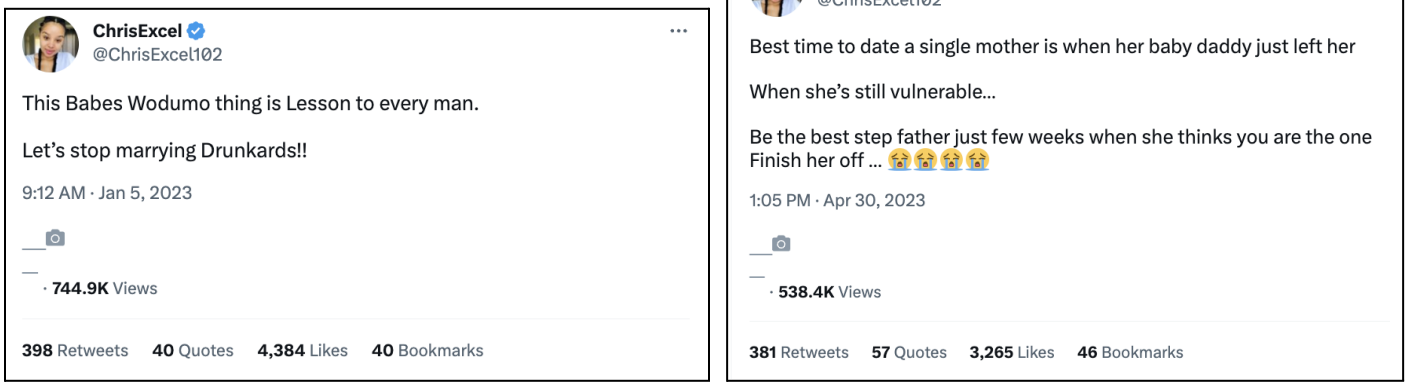

Figure 4: Example tweets from the @ChrisExcel102 account

Another account which has previously come under the CABC radar is @PSAFLIVE. With almost 40k followers since joining Twitter last year, this account has positioned itself as a media and news company. The account is also linked to the PSAFLIVE website, an online platform that produces articles centred around alleged crimes committed by immigrants. Examples of tweets from this account can be found in Figure 5.

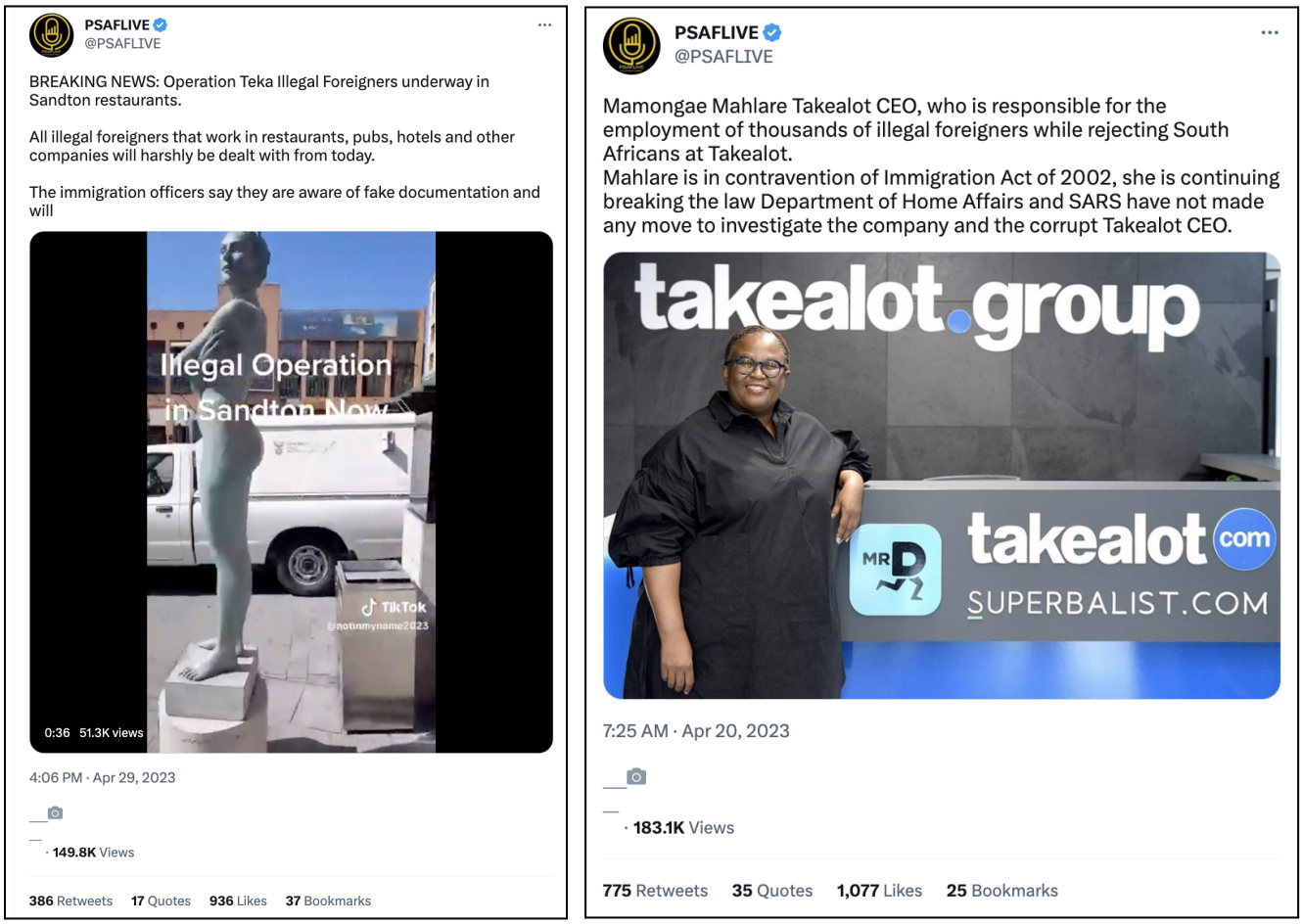

Figure 5: Example tweets from the PSAFLIVE account

An interesting feature of the PSAFLIVE website is that, unlike clickbait news sites that generate income by allowing targeted advertising from platforms like Google Ads, this self-proclaimed news site doesn't appear to use the empty space on their site for the positioning of such adverts. This means that the purpose of this website is not to generate revenue through advertising, raising the question of how this site and its online accounts are funded.

Perhaps it is through the sale of merchandise, as we find on the Instagram account (see Figure 6), further bringing into question the true intention of the person or group of people behind the creation of the PSAF profiles and website. Notwithstanding, PSAF clearly aligns with the emerging - primarily anti-immigrant - rhetoric that has accompanied the rise of political parties and actors in recent history.

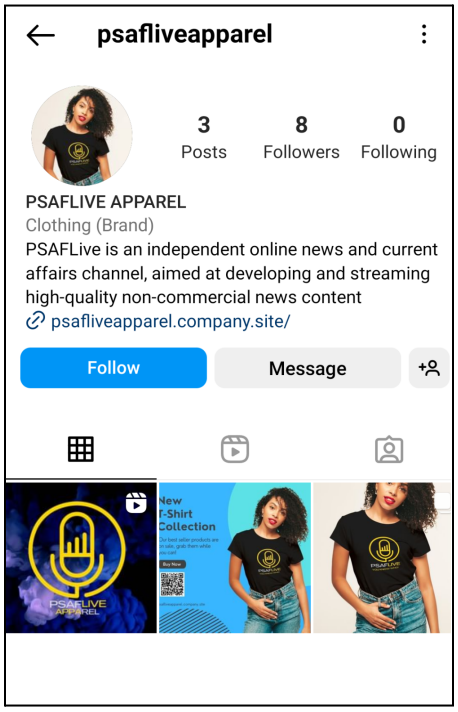

Figure 6: PSAFlive Instagram account.

Further examining the PSAF website, we find a set of terms and conditions that has missing information and tries to distance itself from any third party information on the site, perhaps in an effort to try to distance themselves from the usage of the content that they post (Figure 7).

However, the purpose and intention behind the creation of this site, much like their Twitter and Instagram profiles, are difficult to understand on a brief review given the haphazard nature in which the site is constructed.

Figure 7: Poorly written Ts and Cs on the PSAFLive website.

Returning our focus to Twitter, a closer look at the top hashtags used by the PSAFLive account latest 2000 tweets reveals a central focus on issues of immigration, as well as other groups that have spread the anti-foreign national rhetoric online, including Operation Dudula and #PutSouthAfricaFirst (see Figure. 8).

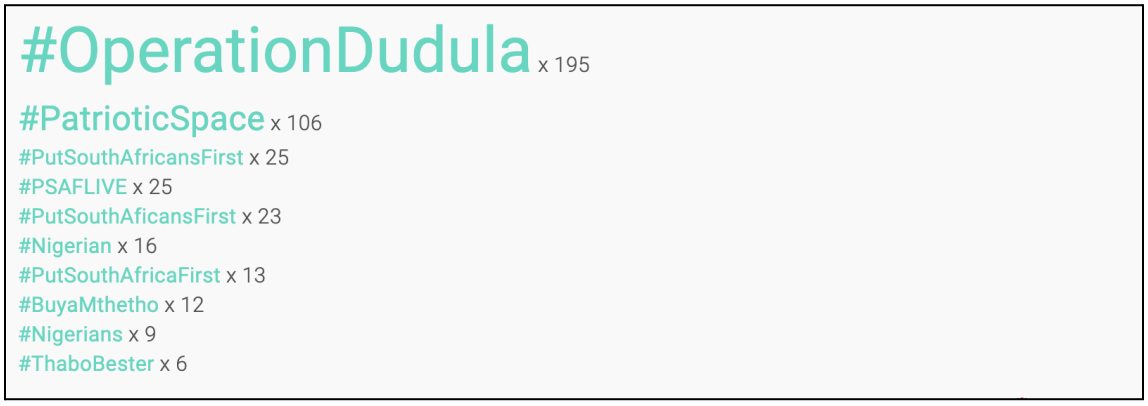

Figure 8: @PSAFLive most used hashtags in the last 2000 tweets

The credibility of some of the posts also comes into question, as is the case in Figure 9 image below, where the account alleged that Nigerians were being executed publicly in Saudi Arabia, after they were arrested for drug trafficking and trade.

Figure 9: Example tweets from the @PSAFLive account

A Tineye[22] reverse image search of the still image from the video of the post - which the account was "unable to tweet…due to its graphic nature" - revealed that the image had been online as early as 2014.

The account also makes unsubstantiated claims about alleged crimes committed by foreign nationals, with no evidence to back them up. These tactics have been identified by CABC researchers in previous reports, dating back as far as the @uLerato_Pillay xenophobic network[23].

Conclusion

Twitter's new blue verification process may have yielded more positive outcomes than it has been credited for. The inclusion of a credit card during the subscription process and further requiring a registered South African mobile number means that the complete anonymity of the person or people behind the accounts should no longer present a challenge to law enforcement authorities. When tracking down the purveyors of disinformation and hate speech, such authorities may now have recourse to request the personal details connected to that credit card or the phone number that was used to create an account. As an intervention, this is promising because it could imply a clamp down on content that has the ability to impact social cohesion in South Africa.

[22] https://tineye.com/about

[23] https://www.dailymaverick.co.za/article/2020-08-18-ulerato_pillay-how-the-xenophobic-network-around-putsouthafricafirst-was-born-and-then-metastasised/

However, it is important to emphasise that for the moment Twitter may only reveal the identity of a user if the account is found to be posting content that is in the public interest. Proving that the content posted by an account meets this requirement will undoubtedly be very challenging. Based on our learnings from the handle @ChrisExcel above, questions arise on whether the policy around parody, fan or commentary accounts needs to be tightened, especially in cases where a fake identity is used to spread derogatory content - about an individual or a group of people online.

Importantly, from a longitudinal viewpoint, a measurement of the impact of this intervention is necessary to understand if it really is effective in the broader fight against online mis- and disinformation sharing and polarisation of the public around key topics of concern. Measuring the impact of these interventions and comparing this efficacy to other interventions that have been put in place on social media networks is of importance to both private and public sector actors as it has the potential to inform policy making.

** Our special thanks to Cathleen Powell for offering her legal expertise and time to review this report.* *Cathleen Powell is an **Associate Professor in Public Law at the University of Cape Town**, holding a BA and LLB from the University of Cape Town, South Africa; an LL. M. from the Humboldt University in Berlin, Germany; and an SJD from the University of Toronto in Canada.*