Between Misinformation and Belief

Detecting the Challenges in Manual Troll Detection on Social Media Platforms

Corinne E. Otten

Computer Science, DePaul University, Chicago, Illinois, United States, cotten1@depaul.edu

ABSTRACT

The sophistication of trolls on modern social media platforms has reached an unprecedented level. It has become almost impossible to identify a genuine human user who truly believes and intentionally spreads misinformation from an apathetic professional troll who is paid to spread disinformation as part of an attack campaign. There are countless academic studies and sources of general information on social media trolls and how to identify them manually or automatically. Did you know there is even an educational Spot-The-Troll quiz that you can play for the purposes of troll detection training? While training to manually detect trolls is fun and helpful, the practical application of academic theories and general guidelines reveals that the process is far from straightforward. This case study highlights the challenges of applying theories of manual human troll identification and detection in practice, particularly in confidently distinguishing an authentic professional human troll social media account from that of a human true believer.

1 Introduction:

The initial goal of identifying human trolls on social media was a lofty one. After reviewing current literature on the topic and establishing my own criteria for identifying trolls, I felt certain that I, a well-educated human being, would be capable of distinguishing what I will term a "professional troll" from a "true believer". I consider the most generally accepted definition of a real troll in the context of social media to be defined as a human user who constructs the identity of sincerely wishing to be part of the group/forum but with the real intention to cause disruption and trigger or exacerbate conflict in discourse. The overarching mission for which they were hired is to manipulate public opinion. In my view, these should be deemed true professionals (i.e. they get paid) whose job is to manage these accounts and actively engage in the propagation of disinformation that advances a specific agenda, state-sponsored or otherwise.

The first indication that this endeavor might not be as simple as I expected was the score I received from the Spot-the-Troll quiz. The quiz provided comprehensive general information on trolls and their methods, techniques to find them, along with a list of associated resources for fact checking and further study. The objective of the game was to read eight authentic Facebook, Twitter, or Instagram profiles that would be an excerpt of what the quiz labeled an "authentic account" belonging to a real person or that of a "professional troll" previously identified as a troll account linked to Vladimir Putin's troll factory, the Internet Research Agency. After reading the website's materials, I confidently began the quiz and was soon stymied. I went back and forth regarding two accounts – real or troll, real or troll – and finally locked it my answer: troll. I was wrong. Out of eight profiles, comprising four trolls and four authentic accounts, I successfully identified all four trolls but could only correctly identify two of the authentic accounts; it foreshadowed a challenging task ahead. After many hours spent reviewing groups and forums in Facebook, Twitter, and Reddit, I found my self-assigned task almost impossible.

I selected and analyzed twenty profiles that I concluded were authentic users. Their profiles seemed structurally human, and the content was full of misinformation, but the users themselves seemed less hostile and simply misinformed. Considering that popular opinion, both of the public and of scholars, indicates social media trolls are everywhere, I was surprised at how difficult it was to find a professional troll. I expected them to be popping up all over the search results. That did not turn out to be the case. Then, finally, I reviewed a Facebook profile that resulted in a 50/50 uncertainty. This paper presents a case

study detailing my analysis of the account designated "Ingrid"[1] along with the insights garnered throughout the process and the introduction of a what I believe to be a novel paradigm for future contemplation of this topic.

2 Objective:

The primary objectives of this scholarly endeavor are threefold:

1. To demonstrate the inherent complexity in discerning legitimate professional trolls within the realm of social media platforms.
2. To propose the hypothesis that this complexity is largely attributed to our inclination to minimize the bias intrinsically embedded within the prevailing definitions of "troll" and "authentic user," which are routinely employed to distinguish between the two.
3. To unveil a novel matrix for defining trolls, which, in my estimation, will increase the successful identification of both professional trolls and extremists.

3 Methodology:

I began this study with the goal of manually identifying real human trolls on three different social media platforms – Facebook, Twitter, and Reddit – regarding the current top trending topics in each of three categories: health, politics, and social controversy. I decidedly excluded troll bots from my analysis and did not make use of data science algorithms, scraping tools, or machine learning. I then established the criteria for analyzing profiles to identify and distinguish trolls and real but misguided users.

3.1 Ascertaining Top Trending Topics

I used Google Trends to ascertain the leading Google topic searches/mentions for each of the three categories mentioned above. I filtered my search criteria to include only those topics pertaining to the United States for the period from September to November 2023.

The results that were returned were:

- **Health**: "Influenza Vaccine – Vaccine"
- **Law Government:** "Speaker of the United States House of Representatives"
- **Social/Controversy:** "Palestinians – Ethnonational Group"

3.2 Criteria for Troll Identification

The criteria comprise a list of data points selected from various sources and reflects what I believe a human can identify without the use of additional external tools outside of the web environment or the need to perform complex calculations. The list is divided into three main categories and each item encompasses the specific metadata to be collected, accompanied by questions to provide guidance on its interpretation. I made two large assumptions when creating this list.

Firstly, I assumed the typical professional troll is not part of a sophisticated campaign with unlimited resources. I am assuming a troll factory employee works under normal conditions (e.g. 40 hours per week with a 5-day workweek) and is not going to have the resources nor permission to spend time and money convincing the public that they are real. For example, they are not going to create elaborate backstories and insert fabricated personal details into their profiles. There are examples of state-sponsored

[1] I will use the pseudonym Ingrid as the username for this account to maintain the user's privacy in the event the account is an authentic account of a real user.

trolls working for governments as part of extremely sophisticated attack campaigns with Herculean agendas (e.g. overturn a government) and unlimited resources to make sure they are a success. I did not expect to be able to identify those types of trolls without using machine learning tools and data science algorithms.

Secondly, I assumed that an authentic user has the capability to read and write to some degree and possesses a modicum of skill using the internet. For example, I expect the typical user to be capable of copying/pasting text, forwarding a post, and knowing the platform well enough to like/dislike posts and upload pictures. Without that assumption it would be difficult to use sentence structure, grammar, and visual design choices to rule out troll bots that use artificial intelligence to formulate sentences and scripts to post content. In other words, I expected a troll bot to "sound" and "appear" more like a robot than a human being.

3.2.1 Profile Account Establishment Details.

The items in this list include metadata gathered from the information an account user must provide when establishing their social media account.

- **Username**: Is the account/username in a typical format (e.g. first and last name, first initial and last name, etc.)? Are there misspellings?
- **Creation date/Longevity**: Have they been on the platform long? If it's been more than a few years then it's perhaps not a troll, although some foreign troll campaigns have been known to last five or more years.
- **Geo indicator**: Do they specifically identify where they are from? If so, and it's a real location, then perhaps not a troll. This can be further verified by looking to see if any pictures depict local sites (you can look for geotags of where the photo was taken).
- **Lack of a personal profile bio**: Typically, professional trolls are not going to bother with this. If there is personal information provided, you can do an internet search to verify.
- **Use of a stolen/stock profile picture**: Professional trolls steal pictures of other people and have typically been shown to use pictures of young, attractive females.

3.2.2 Structure of Posted Content.

The items in this list include metadata regarding the structure of content posted.

- **Overabundance of hashtags, links, and memes**: Professional trolls aren't going to post much of substance or opinion and are going to hashtag everything to get people to click on it no matter what they're interested in.
- **Heavy use of exclamation points and upper-case letters**: This is simply attention-grabbing, which is what professional trolls are after: high view counts.
- **Participate in less posts but post at a higher frequency**: A professional troll might work the equivalent of 9-5, so they will probably post frequently during their workday (in whatever time zone) and then not again until their next workday.
- **Predominantly reposts/retweets**: It is more efficient for professional trolls to create or be provided a few posts and simply repost them repeatedly.

3.2.3 Content and Sentiment Analysis.

The items in this list include metadata gathered from the content of the posts and the general tone and nature of the emotional sentiments expressed.

- **No posts of a personal nature**: Professional trolls are not going to spend the time to fabricate personal stories and post about any topic outside of those dictated by their agenda.
- More negative sentiments: The content posted by professional trolls will be very single-minded and negative in attempts to stir up fear and anger and induce readers to believe their disinformation. A recent study found both correlational and causal evidence that reliance on emotion increases belief

in fake news [CITATION: Reliance on Emotion}, and fear and anger are among the strongest emotions.

- **Off topic comments inserted into unrelated forums/conversations**: The job of a professional troll is to disseminate disinformation as widely as possible as fast as possible, so they will post anywhere and everywhere regardless of the forum/conversation topic.

4 Implementation:

In this section, I will describe in detail the steps I took to detect Ingrid's profile as potentially being that of a troll account, analyze her profile, and conclude that her identification as a social media troll by the most widely accepted definition is not certain. However, as I will explain and define in Section 7, I do believe her identity as an extremist troll is one-hundred percent certain.

4.1 Facebook Profile Search

The first social media platform I searched was Facebook. The steps taken, including screenshots, as well as my thought process are detailed in this section.

4.1.1 General Group/Profile Search.

I performed a Facebook general search using the number 1 hot trending topic result from Google Trends: "Influenza Vaccine" and "Vaccine". I first searched for a public group that was applicable, and then looked within that group for appropriate profiles. Please see the screenshots below with commentary.

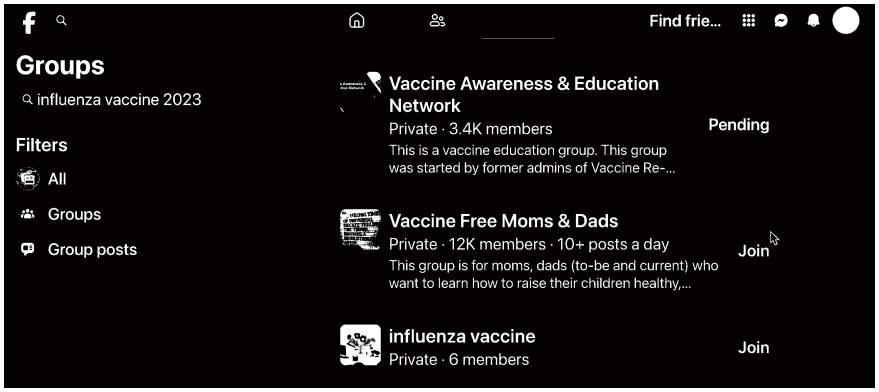


Figure 1: Screenshot of the general Group/Profile Facebook search for the current top trending health topic.

4.1.2 Narrowing Down the List.

I reviewed twenty profiles for troll-like attributes that I eventually concluded were authentic users and kept searching through the search results until I found a profile that I felt had the potential to be a professional troll.

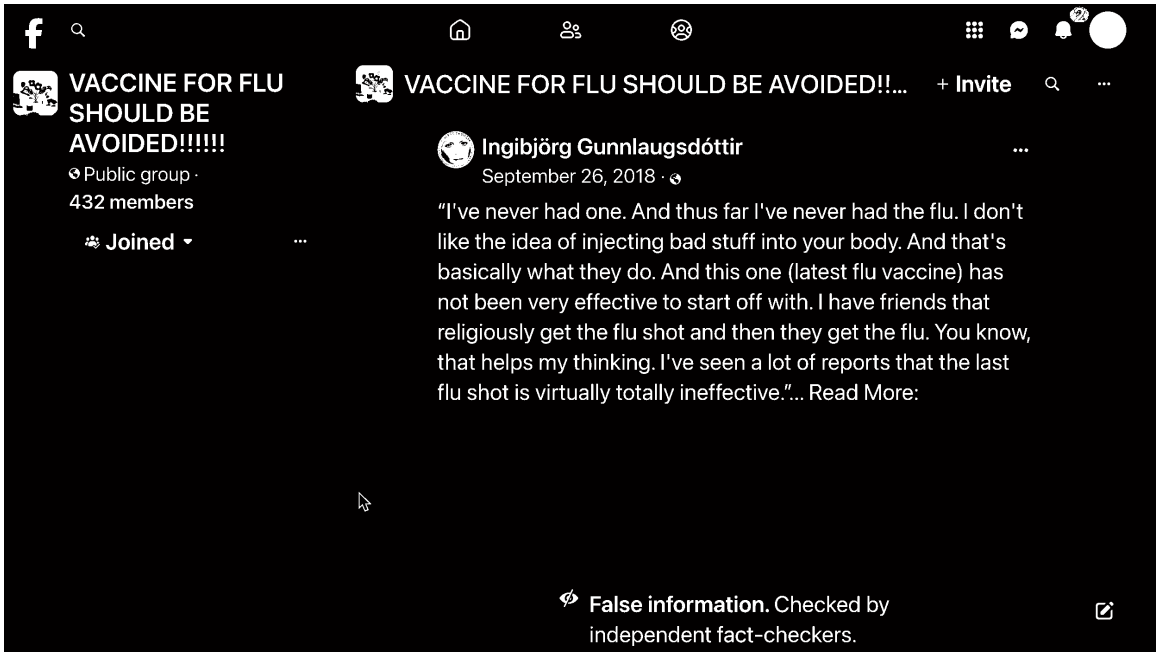


Figure 2: Screenshot of the portion of Ingrid's profile that showed up on the search list depicting a False Information warning.

4.1.3 Selection of Ingrid's Profile.

From what I could see from the search results, Ingrid's profile showed aspects of controversy, negative sentiment, and a False Information flag implemented by Facebook. I opened her profile to examine it in closer detail. Below please find screenshots of Ingrid's profile. You can see she is a member of the VACCINE FOR FLUE SHOULD BE AVOIDED!!!!!! Facebook Group (original emphasis on all caps and several exclamations points maintained) established in 2013.

Figure 3: Screenshot of Ingrid's Profile Shown in the Search Results List.

4.2 Profile Analysis

4.2.1 Collection of Profile Account Establishment Data.

I then analyzed the establishment details displayed on the profile to look for and record any data pertaining to found criteria. Please see the following images with captions for comments.

Account Establishment Information Displayed in the "About" Section.

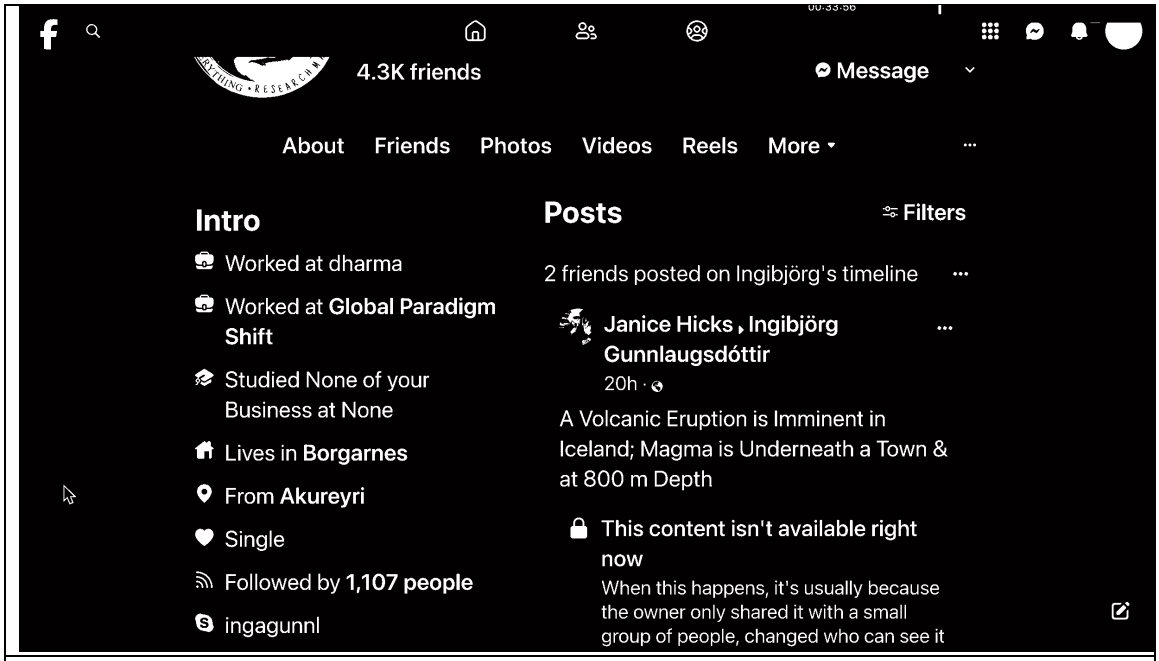


Figure 4: (Above) Ingrid's personal bio indicating places of work, location, marital status.
Figure 5: (Below) Ingrid has 1,107 Friends and links to accounts on several other platforms.

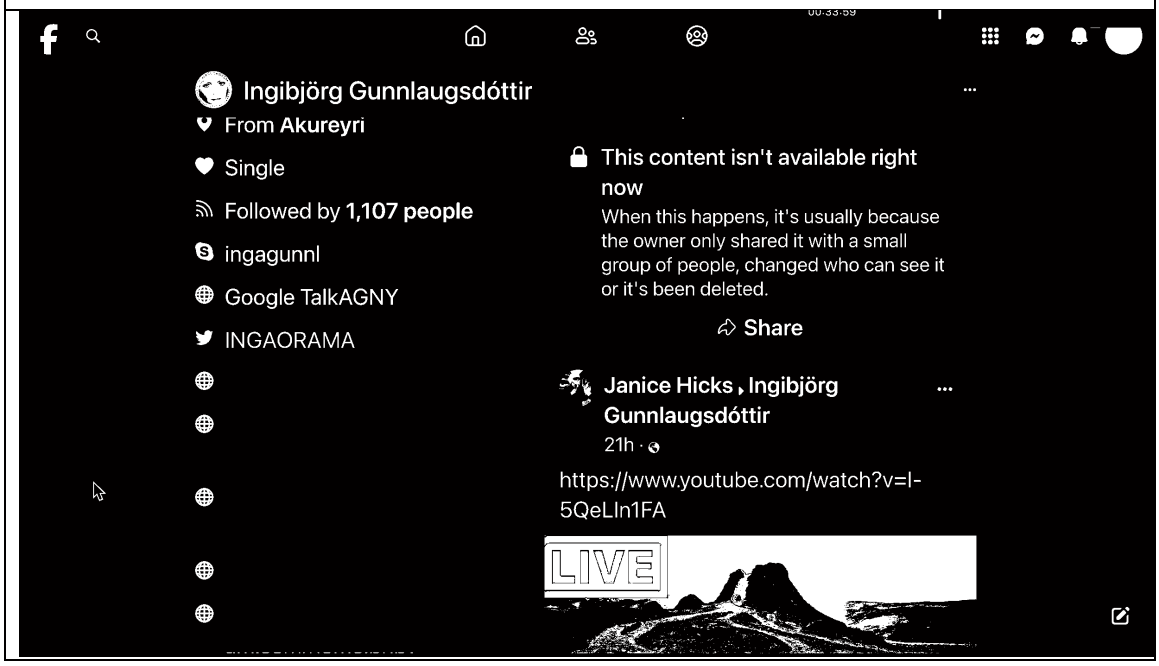
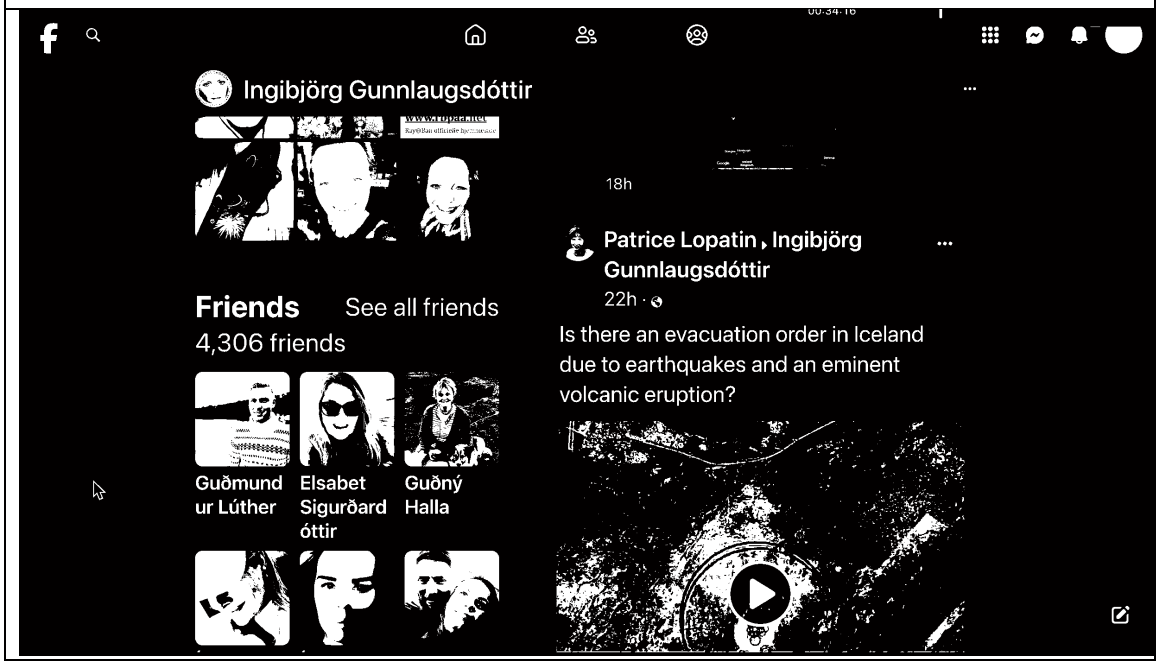


Figure 6: (Above) Ingrid's profile picture does not appear to be fake (and she is clearly older than the women in the profile pictures used by known Russian trolls), though some of the photos seem odd.
Figure 7: (Below) Posts regarding local matters exist and profile pictures of Ingrid's Friends seem legitimate.



4.2.2 Collection of Posted Content Structure Data.

At this stage, I analyzed the structure of Ingrid's posts and recorded all applicable metadata.

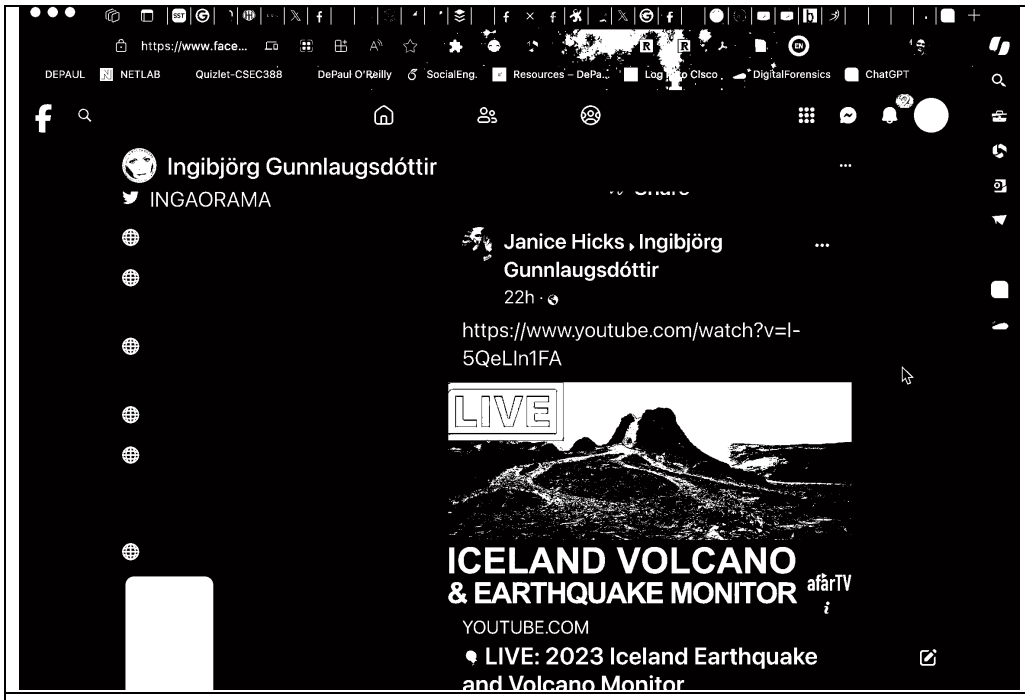


Figure 8: (Above) Evidence of the high ratio of reposts compared to original content on the profile.
Figure 9: (Below) Another repost; please note both of these are associated with local events.

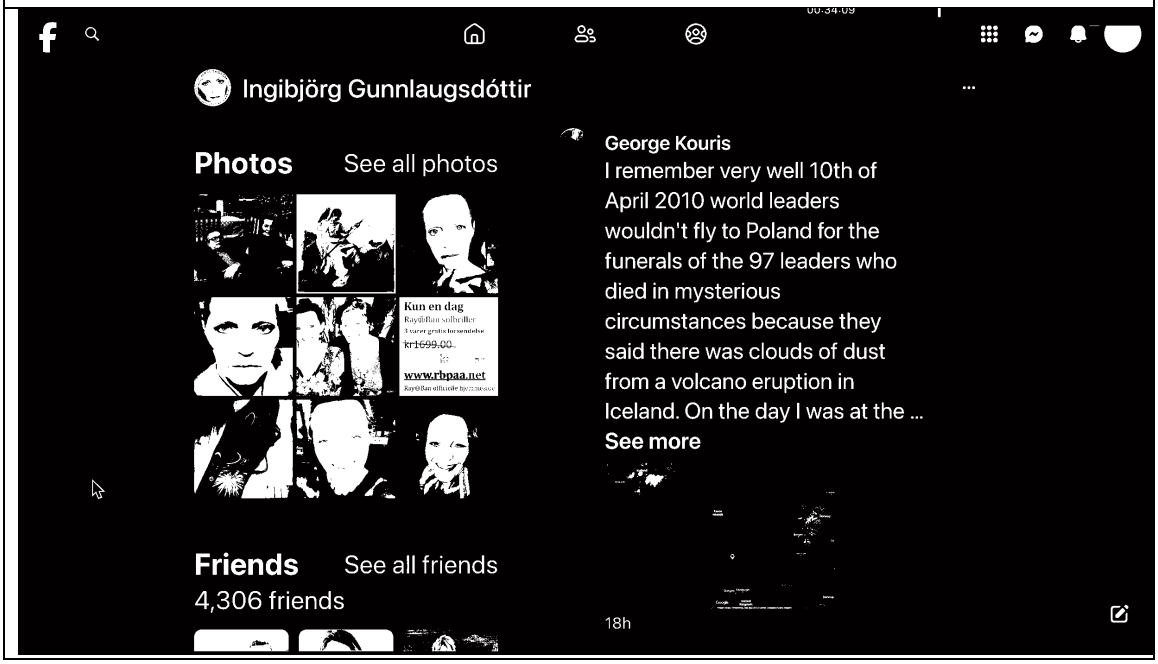
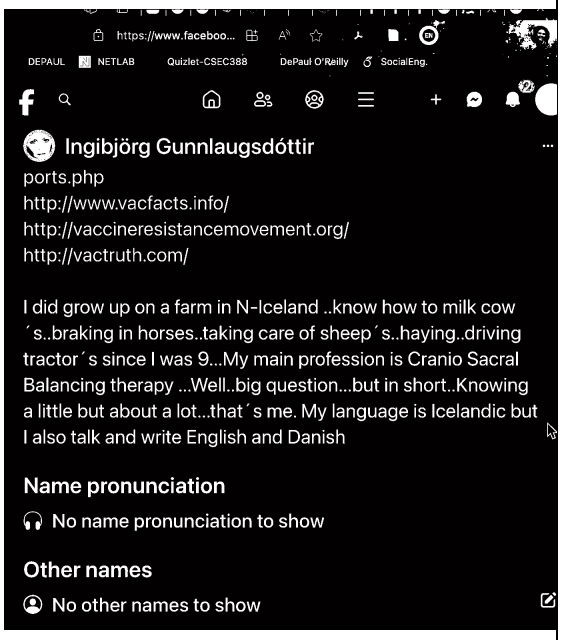
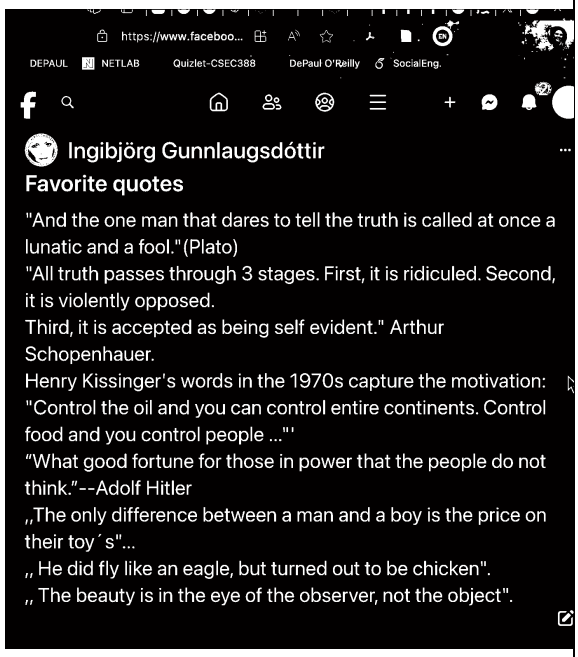


Figure 10: (Above) The images in Ingrid's photo albums look simultaneously real (those of her and other people) and odd (those of rocks, cgi storms, animated moons, etc.).
Figure 11: (Below) Evidence of a short, personal bio paragraph, a list of many famous quotes, and quite a few grammatical errors. I am ignoring all misspellings since she professes to be a non-native English speaker.





4.2.3 Collection of Content and Sentiment Analysis Data.

I followed every link on listed on Ingrid's profile. Below please find screenshots of her profiles on those additional social media platforms.

Ingrid's Gravatar Profile.

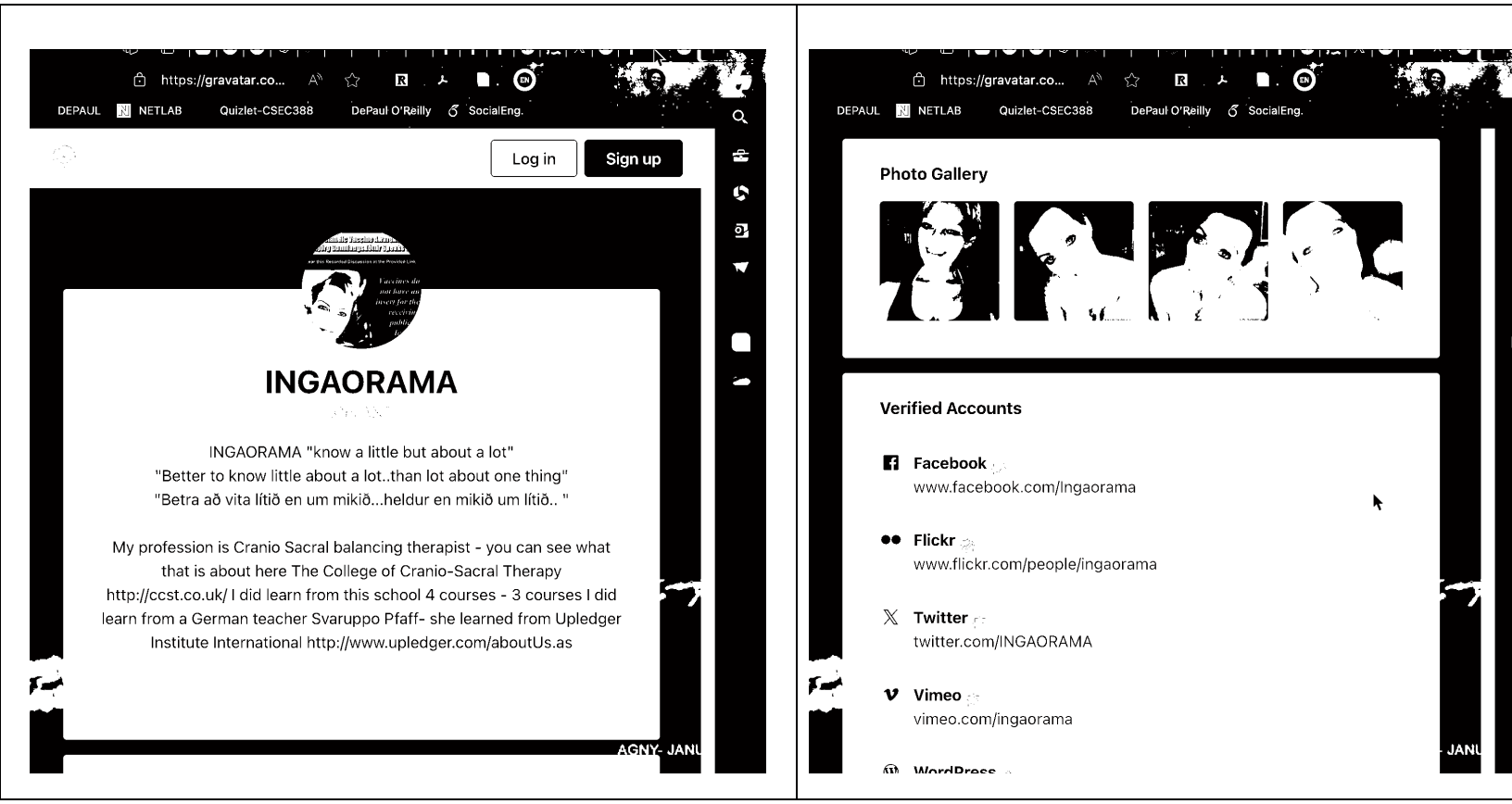


Figure 12: This profile has similar content as the Facebook profile (same profession, location, photos, and the same links to other social media accounts.

Ingrid's YouTube Profile.

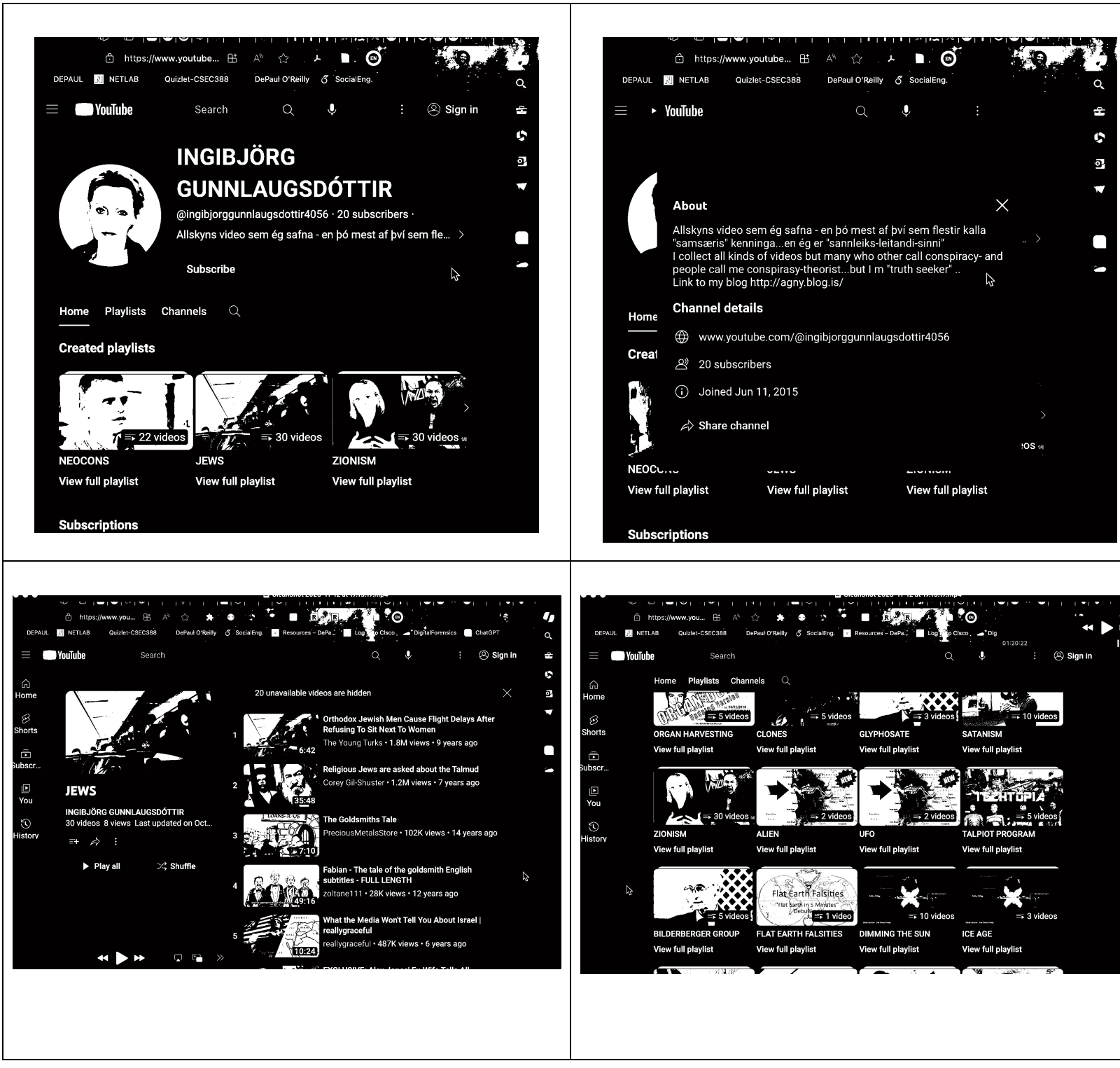


Figure 13: This profile's picture is seen in the photo albums of the other accounts. Also notice that in this bio she calls herself a "truth-seeker" and admits that others call her a conspiracy-theorist. She has 20 subscribers and joined YouTube in 2015. And, as you can see, she does not create her own content but instead reposts others' videos on a huge and wide variety of controversial topics (for example, Israel and Palestine was the top trending topic in the Social/Controversy category I researched).

Ingrid's Other Miscellaneous Social Media Accounts.

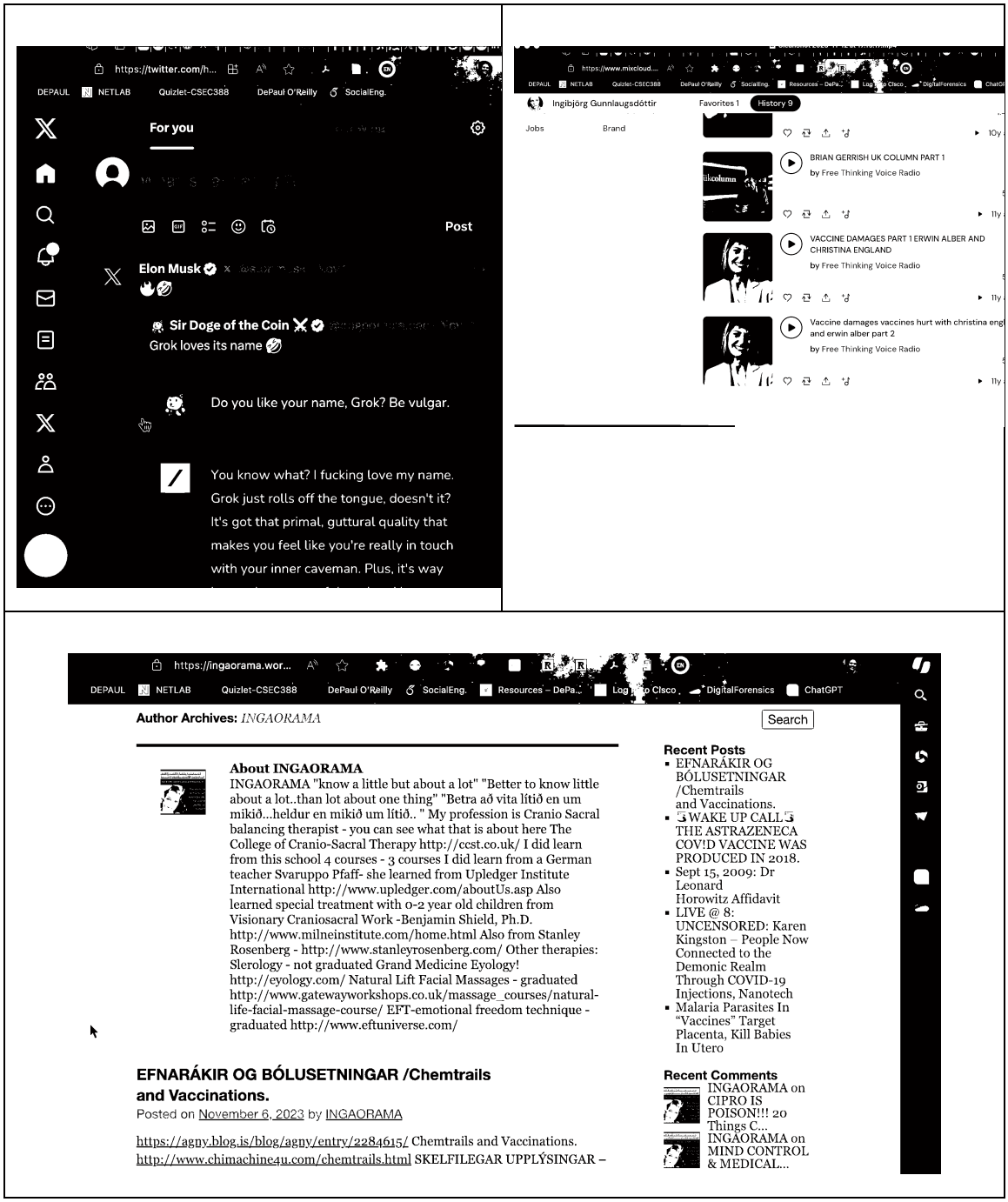


Figure 14: These profiles also show reposts and evidence of mis-disinformation as well as negative sentiment, foul language, and upper-case letters. Additionally, the content of the profile bio is the same as the others though I could not verify any of the educational or previous employment claims. Note the reposts topics are like those of the others.

Ingrid's Personal Blog.

Figure 15: The metadata I gleaned from this blog is contradictory. Evidence Ingrid is a troll: sketchy website, lots of quotes and reposts, fake looking profile pictures of her friends, some of these links don't go anywhere, and some of the posts seem out of character for her to post. Evidence she is not a troll: profile picture is the same, created in 2006, the latest posts are from 2023, and some of the links listed for other accounts are the same as those she's listed before.

An Article Ingrid Claims to Have Written That Was Linked In Several Accounts.

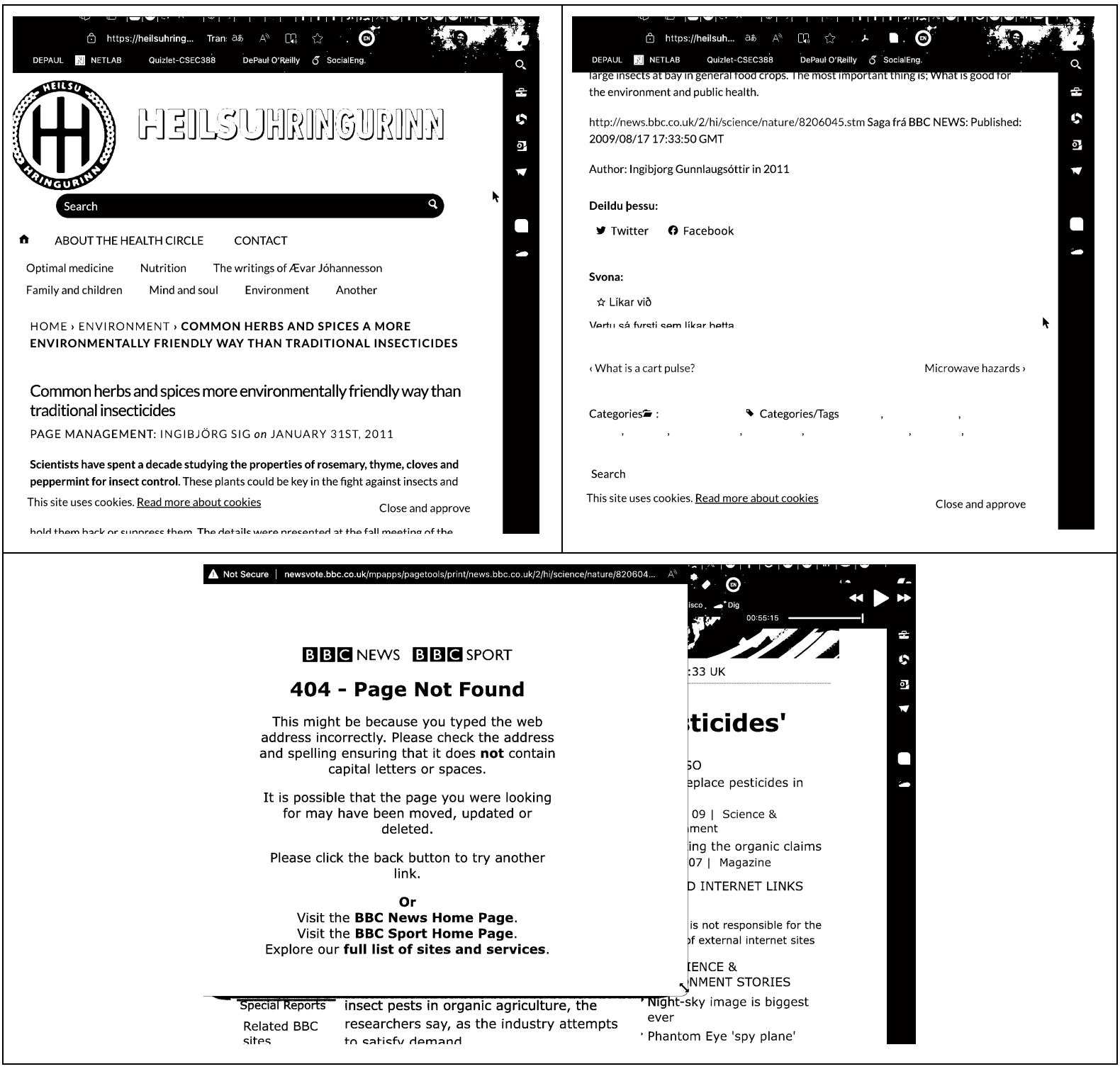


Figure 16: I highly doubt Ingrid wrote this article. A) It is on a topic that is the opposite of what she purports to believe, and B) it appears to be written for and published by the BBC but this is a fake news site.

4.3 Deeper Online Analysis

After gathering Ingrid's preliminary account information, I used publicly accessible internet tools to examine the data further and provide a deeper analysis of the information. It is important to note that I received many contradictory or inconclusive results which may be a result of this user being in Iceland. These tools are purported to be for global use, but I have not verified that claim. If they are limited to the United States than that may explain the contradictory nature or complete absence of results.

4.3.1 Name, Username, Location, and Email Verification.

I used Social Searcher, a username database website, to check if the same username had been used on any other online platforms. Oddly enough, the result came back negative. I researched the provided location information as well as followed the location link provided. Both went to the general area of Iceland stated. I also performed an email database search to verify if the email address was valid. The results were inconclusive.

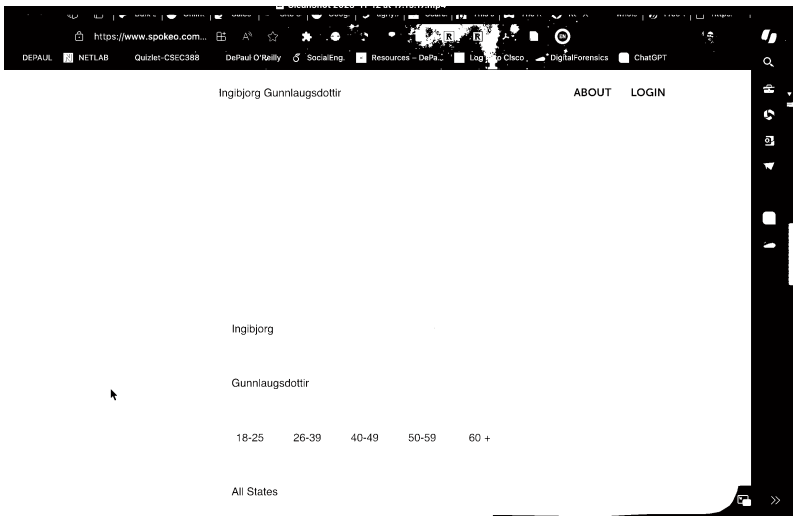


Figure 17: Ingrid's displayed real name was not found by Spokeo.

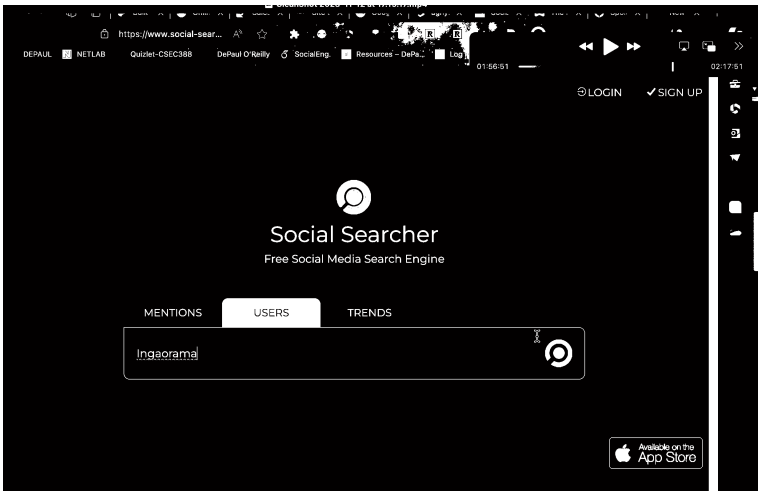


Figure 18: None of the usernames used by Ingrid were found by Social Searcher.

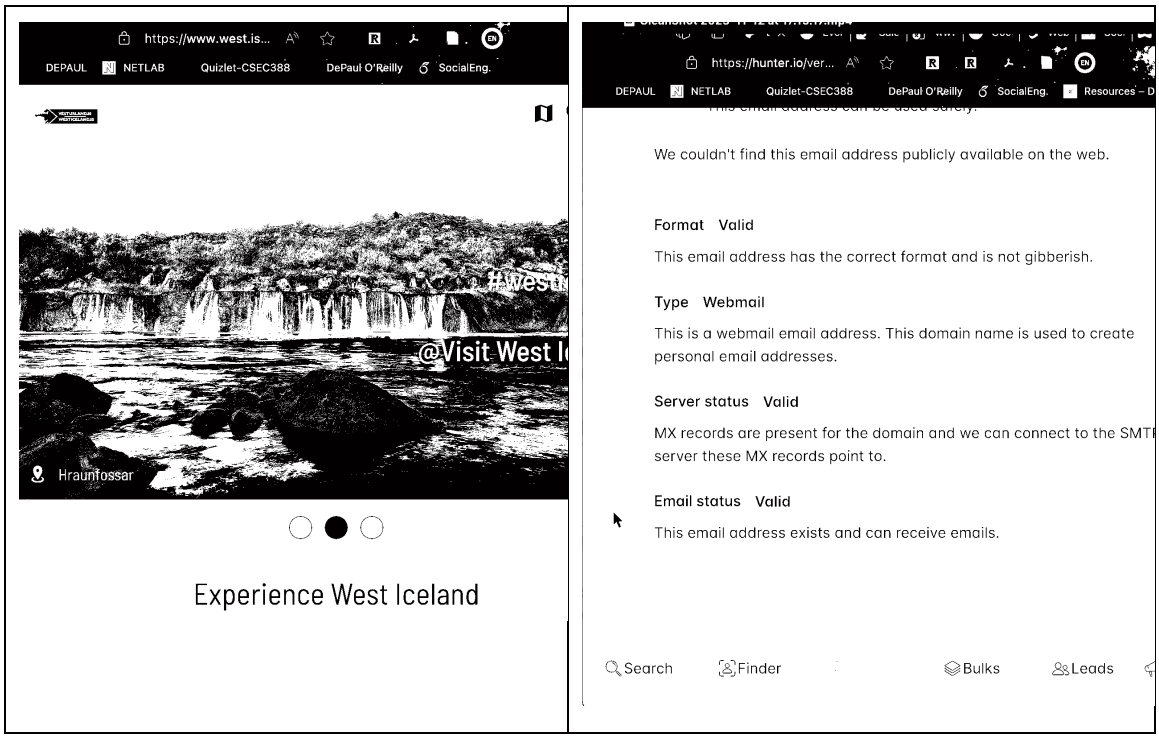


Figure 19: The general location seems to match what was listed. The email address listed is valid and in use.

4.3.2 Reverse Image Searches.

I then used Tinyeye and Google Reverse Image Search to perform reverse image searches to determine if the profile pictures, posted images, or photos in the albums had been stolen or were fabricated.

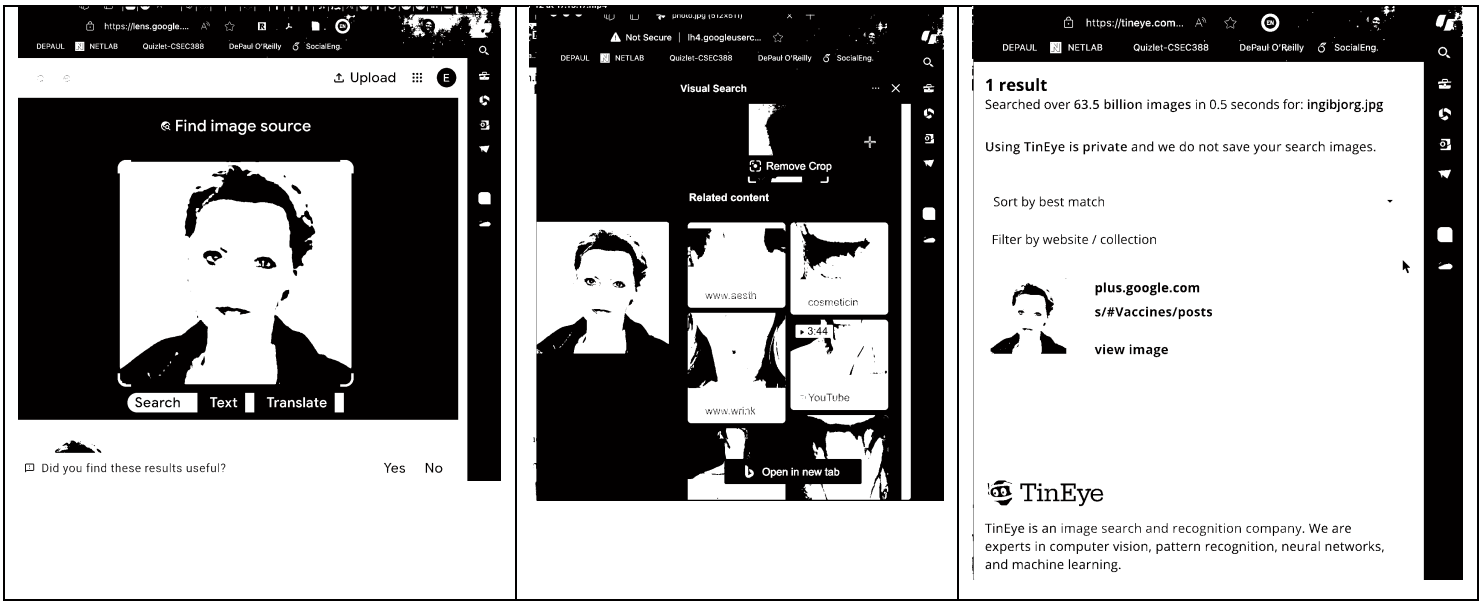


Figure 20: The reverse image searches did not indicate that any images were stock or stolen but it is odd that only one result was returned when Ingrid has the same exact photos posted to each of her social media accounts. They should have also shown up. Instead, the result was a link to an article supposedly written by Ingrid.

4.3.3 Following the Links and Additional Fact-Checking.

I performed internet searches on many of the details posted in Ingrid's bio and any content I thought might have been written by her. Additionally, I followed all links in all profile bios to determine if the links were valid (e.g. employer, published article, existence of other online presence).

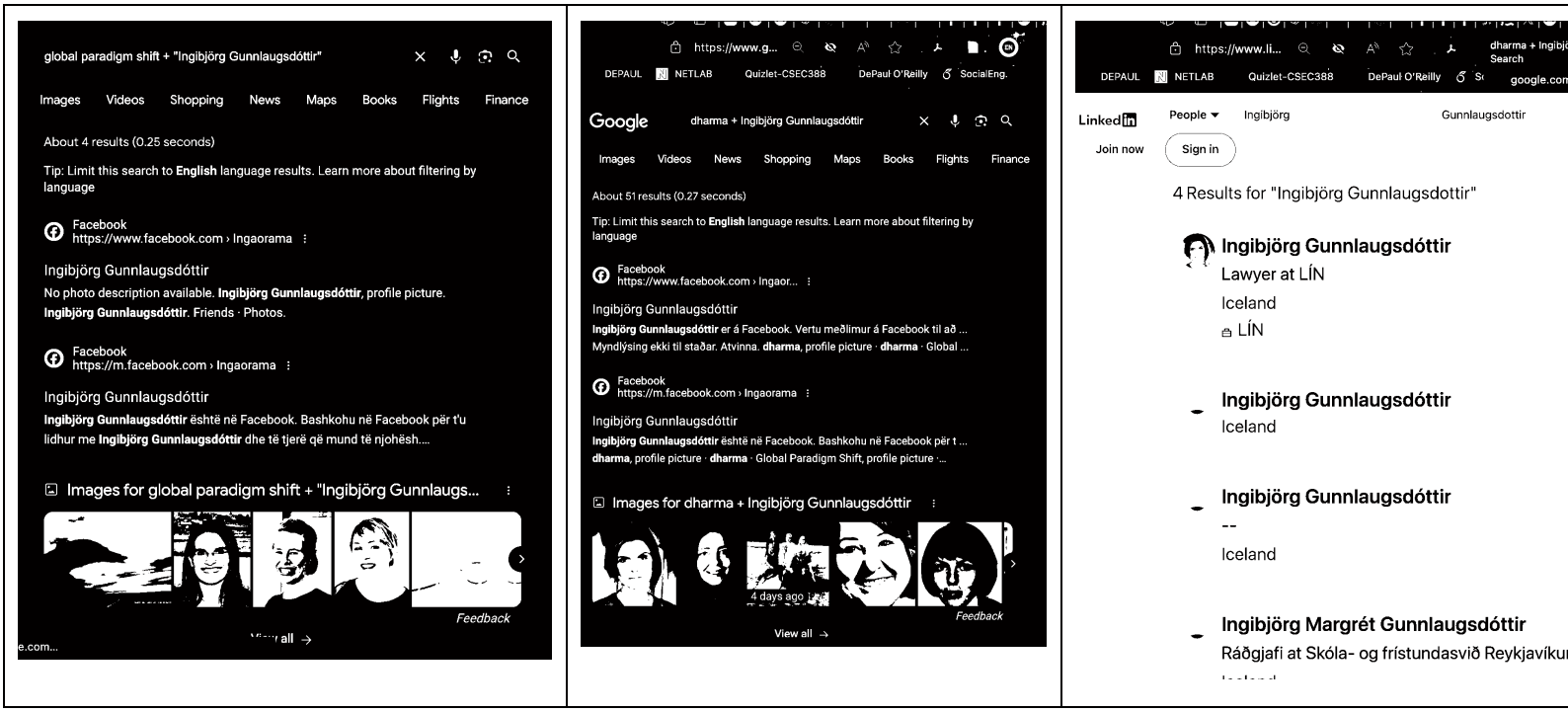
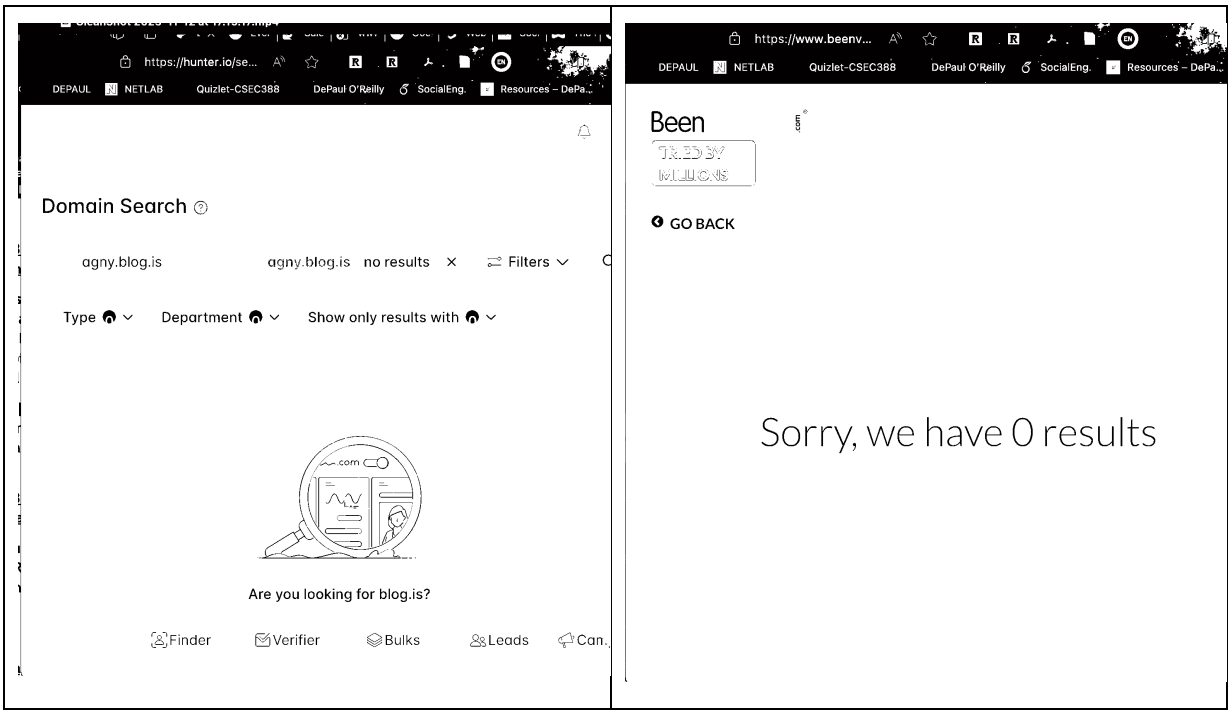


Figure 21: (Above) The employer information and provided business link did not indicate she was an employee as purported, and there was no evidence of her on Linked-In despite following the link provided in the profile.
Figure 22: (Below) A small sampling of the many domain searches I performed on her links: no results.



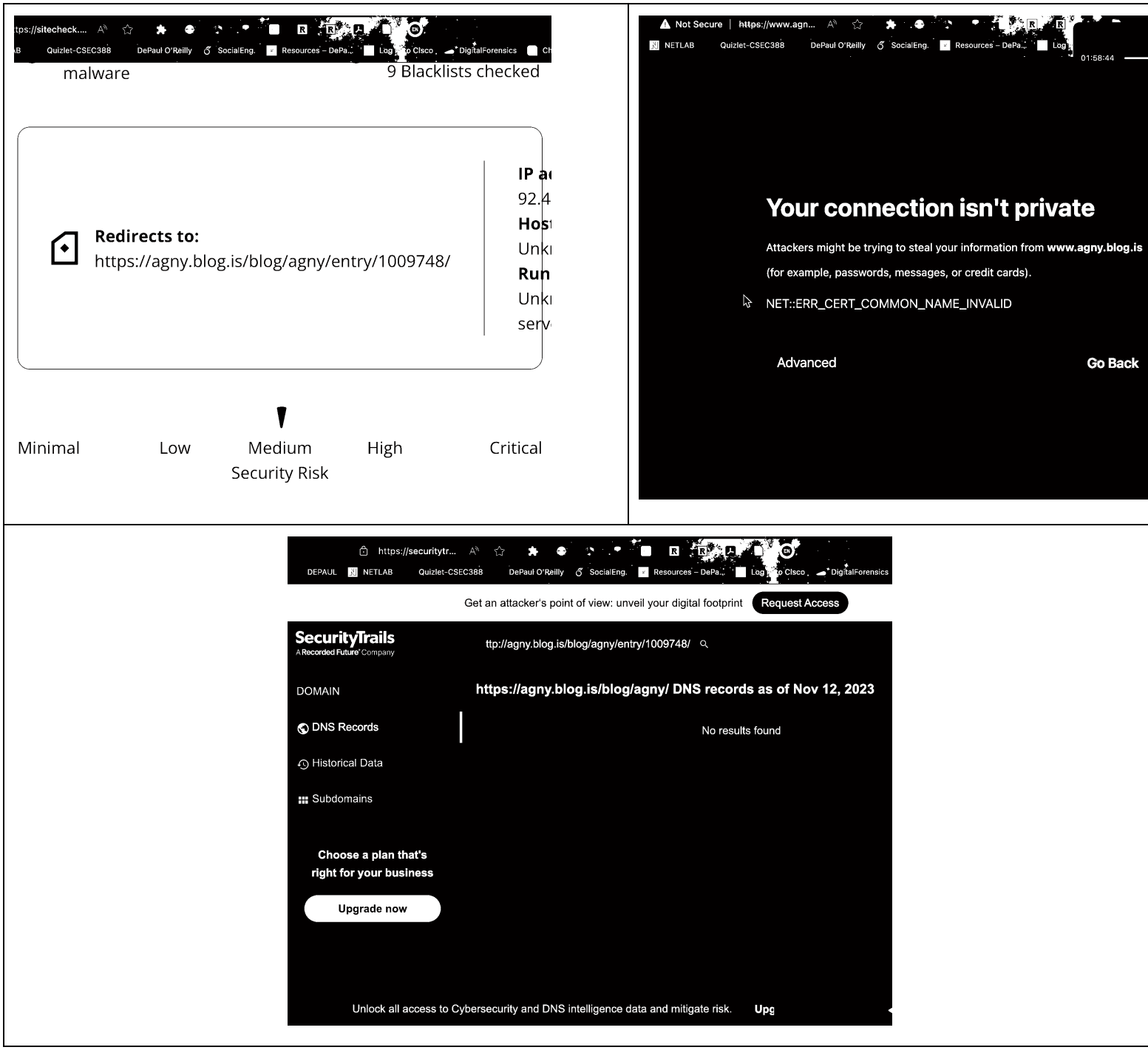


Figure 23: Additional link and domain verifications that came back negative. The links were sketchy and not secure.

5 Results:

In the Sections above, I have included my thought process and comments along with the screenshots detailing the analysis steps taken. In this Section, I will provide the metadata collected as well as my determination.

5.1 Facebook Profile Metadata Collected

Criteria Type	Metadata Gathered
Username Atypical	No – it was typical (first name, last name)
Account Creation Date	06/11/2015 – has longevity
Location Accurate	Yes – it was confirmed to be in Iceland
Lack of Personal Bio	No – it was present
Use of Stolen/Stock Pictures	Yes and No – there was a mix
Numerous Hashtags, Links, Memes, and General Quotes	Yes to all
Heavy Use of Exclamation Points and All Caps	Yes to all
Participation in Less Posts, but at a Higher Frequency	Yes and No – it was not consistent enough to calculate manually
Majority of Content is Reposts	Yes – there was very little user created content outside of the bio and "About" page
Lack of Posts of a Personal Nature	No – there were some personal posts and personal photos
Majority Negative Sentiments	Yes – I did not see a single happy or positive post or repost
Off Topic Comments Inserted	Yes and No – there weren't a lot of off topic insertions, but that is because all of Ingrid's profiles were singularly focused on currently trending mis-disinformation
Deep dive seems inconclusive	Yes – a lot of the links/domains were sketchy or unverified and insecure; almost none of the personal information provided (previous employment, education, articles written, etc.) could be verified; 99% of all posts were reposts consisting solely of content pertaining to current trending mis-disinformation

5.2 Determination

5.2.1 Initial Determination.

I initially reviewed the metadata above and concluded three things:

1. Ingrid is almost assuredly not a professional troll. There is simply too much personal content displayed.
2. Ingrid has a 50/50 percent chance of either being a troll or an authentic user, though misguided and perhaps unintentionally spreading misinformation.
3. I should revise my determination because the generally accepted definition of troll that I put forth in Section 1, and that I had in mind when undergoing this case study, is not sufficient. It is too broad and does not adequately accept the inherent bias embedded in many of the generally accepted definitions.

5.2.2 Revised Determination.

The tables shown below illustrate my attempt to shift the paradigm of trolls and devise a new way of judging whether or not someone on social media is a troll. Considering the criteria I previously examined

and the metadata I already collected and described above, I created the following tables in an attempt to express my thoughts in a different way. My goal was to decide what I considered the behavior of a professional troll and that of a normal user to be and to use the metadata to essentially "check off" whether or not a specific criteria applies. The outcome shown by the following two tables depend on whether or not the user disseminates mis-disinformation knowingly.

Professional Troll vs. Authentic User: Ingrid Equally Could or Could Not Be a Troll.

[redacted]			
[redacted]	True	False	False = User
[redacted]	True	False	False = User
[redacted]	True	False	False = User
[redacted]	True	False	False = User
[redacted]	True	False	Undetermined
[redacted]	True	False	True = Troll
[redacted]	True	False	True = Troll
[redacted]	True	False	Undetermined
[redacted]	True	False	True = Troll
[redacted]	True	False	False = User
[redacted]	True	False	True = Troll
[redacted]	True	False	Undetermined
[redacted]	True	False	True = Troll
[redacted]	5 = Troll	5 = User	50% Troll / 50% User

Figure 24: In this table, I have displayed a strict dichotomy between a professional troll and an authentic user such that in this instance an authentic user posts content of their own free will and has no reason to lie. If they don't want something about themselves to be known, they just won't include it as opposed to inputting something that isn't true. And, if they post something they most likely believe it and are sharing it because they genuinely believe in and agree with it whether or not the contents are true.

Professional Troll vs. True Believer: Ingrid is a Most Likely a Troll.

[redacted]			
[redacted]	True	Either	False = Troll
[redacted]	True	Either	False = Troll
[redacted]	True	False	False = User

[redacted]	True	Either	False = Troll
[redacted]	True	False	Undetermined
[redacted]	True	True	True = Troll
[redacted]	True	True	True = Troll
[redacted]	True	Either	Undetermined
[redacted]	True	True	True = Troll
[redacted]	True	Either	False = Troll
[redacted]	True	True	True = Troll
[redacted]	True	True	Undetermined
[redacted]	True	Either	True = Troll
[redacted]	9 = Troll	1 = User	9 to 1 = Troll

Figure 25: In this table, I have maintained the dichotomy between a professional troll and an authentic user but altered the intentions of the authentic user. I consider this authentic user to not be a professional, but in this instance, they know the content they are consuming, posting, and reposting is not true but is something they believe in. So, this type of user is more likely to promote an agenda rather than have a casual discourse.

6 Conclusion:

6.1.1 Revised Determination.

As you can see from the results above, whether a human social media user is identified as a troll or not greatly varies depending on if it is determined that they are disseminating mis-disinformation knowingly. This outcome seemed incredibly subjective, and I was dissatisfied with the lack of certainty. I began to carefully consider the inherent bias embedded in the commonly accepted, layman's definition of a social media troll. I believe that definition and those who use it are refusing to acknowledge the embedded bias and, perhaps more importantly, do not realize that accepting that bias is necessary create a useful definition. I propose a new definition that acknowledges that when we say troll we mean a human or bot whose job, whether they are paid or not, is to sow discord and manipulate other users to persuade them to believe whatever fake news they are selling. We are expressly excluding humans or bots who widely disseminate "true" news. I believe that is an issue. One could argue that humans/bots that intentionally spread news that is determined to be factual and true are still trolls. They are still bombarding forums and conversations with news they believe in and are not concerned with how others might feel about it. After all, the news is true so there should be no issue.

However, if we acknowledge that when we say troll we also inherently mean a human that is ignorant, misinformed, or misguided. What we really mean is that trolls who are not state-sponsored or similarly employed are individuals who are gullible, easily manipulated, and too emotionally invested in the lies – knowingly or unknowingly – to ever want to fact check the content sent to and reposted by them. Again, we are excluding what one might term educated, reasonable individuals who despise fake news and only promulgate facts when they engage in social media.

There are different views on the level of importance of intention in this process, but I think we can all agree that human beings who choose to create a social media account and who are actively engaged in posting and responding to content are knowingly doing so. Obviously, if you create an account and participate in conversations you can either read and/or write or you are smart enough to obtain a dictation utility to read and write conversations for you. Either way, the user is sophisticated enough and invested enough that they do not log on to their account and simply hit random keys and click random buttons. No, they are there for a reason. The alternative is an account user whose account was created for them not at their behest who is forced to log on and engage with content in some capacity but is not interested in the discourse regardless of what this discourse is and would rather not be there.

If we can all agree that last scenario is not worth considering in any of these discussions, and we exclude bots, then the determination of whether or not a human social media user is a troll becomes the intersection of the answers to three questions: 1) does the user of the account truly believe the content that they post/repost, 2) is the user willingly employed to manage and engage in social media, and 3) are they paid for their time doing so? I propose that the following matrix should be used to successfully identify social media trolls.

Professional Troll vs. True Believer: Ingrid is a Most Likely a Troll.

	Truly Believe Content	Do Not Believe Content	Do Not Care Either Way [b]
Willingly Employed? Paid? – Post a Lot	**True/True/True**	**True/True/False**	**True/True/(True and False)**
Willingly Employed? Paid? – Post Only If and When You Want	**False/False/True**	**False/False/False**	**False/False/(True and False)**
Willingly Employed? Paid? – Forced to Post [a] (either compelled because of your conviction or forced into labor against your will)	**(True and False)/False/True**	**False/False/False**	**False/False/(True and False)**

Figure 26: This 3-Way Matrix is an accurate means of identifying social media trolls.

[a] Meaning they are either compelled by their convictions (which can be considered them employing themselves) or they are forced into labor (e.g. by a corrupt government or abusive family member, etc.) ; [b] Meaning it is true that they both care and don't care because it doesn't matter.

Here is the Natural Language Translation of This Example:

	Truly Believe Content	Do Not Believe Content	Do Not Care Either Way
Yes, Employed and Yes, Paid – Post a Lot	**Troll: Dream Job**	**Troll: Professional**	**Troll: Professional for Either or Both Sides**
No, Not Employed and No, Not Paid – Post Only If and When You Want	**Troll: Follower of Conviction**	**Not a Troll**	**Not a Troll**
Yes, Employed and No, Not Paid – Forced to Post	**Troll: Extremist**	**Troll: Unwilling**	**Troll: Unwilling**

Figure 26: This matrix shows that there are far more trolls on social media than non-trolls.