Detecting "Fake News" on Facebook

Hannah Eyre Zane Zakraisek

December 8, 2017

The term "Fake News" gained popularity during the United States' 2016 Presidential Election to describe a rapidly spreading phenomena of news articles deliberately spreading false information, often through attention grabbing headlines or headlines that resemble legitimate sources (**guardian**). It became particularly notorious on social media sites and Facebook in particular. Here, the top 20 articles from fake news sites and hyperpartisan blogs garnered more user interaction between August 1st and election day on November 8th than the top 20 articles from a variety of established news sources. These included *The New York Times*, *Washington Post*, *Business Insider*, and Fox News (**buzzfeed**).

FactCheck.org, part of the Annberg Public Policy Center at the University of Pennsylvania, breaks down how an individual can identify fake news into eight parts (**factcheck**):

1. Consider whether the source is credible.
2. Read beyond the headline.
3. Check whether the author is credible or real.
4. Check whether the article is recent.
5. Check whether the article is a joke or satire.
6. Consider your own biases and how they affect your judgment.
7. Check supporting sources (if any) and make sure they abide by the same rules.
8. Ask experts or fact-checking sites.

Rapidly spreading fake news articles can have a range of consequences. In particular, one instance dubbed as 'Pizzagate" ended with a gunman attacking a Washington, DC pizza parlor over allegations of a satanic child sex abuse ring centered around John Podesta, Hillary Clinton's 2016 campaign manager (**pizzagate**). Various politicians and government agencies in the United States and internationally have voiced opinions on what qualifies as fake news, and more importantly, what to do about it. Up to this point however, no consensus has been reached. Facebook was initially reluctant to admit there was any problem with fake news on their website. Since then however, Facebook's CEO Mark Zuckerberg has since released a statement describing how

they plan to deal with fake news in the future, including renaming the phenomenon "false news" (**zuckerberg**).

As an increasingly global and hotly contested issue, we would like to to explore what responsibility Facebook has in regards to these eight points. We will discuss whether they have a responsibility to develop tools to detect fake news based off these guidelines and, if these tools exist, whether they should be used to remove content from the site. We will be using two particular examples here; the previous example of Pizzagate, and the recent Las Vegas shooting. We feel that Pizzagate is a primary example of what happens when fake news reaches a critical interest online, having large followings on Reddit, 4chan, Breitbart and Infowars that actively encouraged "citizen investigation". Additionally, the Las Vegas shooting conspiracies are a group of examples of the rapid spreading of fake news, particularly when real information has not yet surfaced. Here, the website 4chan organized a successful effort to manipulate trending topics across the web and spread misinformation before real information was distributed.

The other primary focus of our analyzation makes use of the utilitarian ethical framework. When analyzing any issue, it's important to define which framework you're working in, along with what premises your arguments are based on. During our analysis of the Facebook's role in moderating fake news, we'll be working under a utilitarian framework. In particular, our primary goal is reducing the negative utility fake news brings about. As we examine role that fake news has played in the events mentioned, it is evident that fake news results in a negative utility for Facebook's users. News on Facebook has the potential to result in positive utility, and like fake news, this is due to the rapid spread and wide range of people reached. We believe access to a variety of news sources is an important part of interacting with social media and results in a greater degree of positive utility than fake news causes negative. Because of this, we believe removing news in general from Facebook would increase negative utility overall. As a result, we find that it is far more important to try to minimize the corresponding negative utility caused by fake news, than removing news from Facebook in general.

As we address each of these eight points mentioned above, the format shall be as follows. We'll first examine the status quo where Facebook has not publicly addressed efforts made to combat fake news and how that is affecting users. Next, we'll examine the outcome if Facebook chose to implement a fake news monitoring system using this technique. Next, we'll weigh the different utilities and select the one that we feel minimizes the negative utility. Finally, we'll also be examining whether or not the method is technically feasible to implement with current technology.

1 Credible sources: Sites and Authors

The first method in combating the spreading of fake news is to check whether or not the article in question comes from a credible source. This involves not just the author, but the site that the

article comes from. To start off, we'll examine the utility if Facebook chose *not to* implement this source credibility check. Although Facebook implements a very light form of censorship, the scenario would be quite similar to how Facebook currently monitors articles, which is very little. While most articles that Facebook news feeds show users are based off topics that they've shown interest in, nearly every Facebook user has probably been shown an article from an underground website that someone on their friends list has liked or shared. If Facebook didn't implement source credibility prioritization, then the effect of one person sharing an article like this could have quite a negative outcome. By one individual sharing the article, it can easily make its way to hundreds of people. In most cases, this has the effect of increasing the amount of negative utility, as most fake news articles are spread with negative motivations in mind.

Now assume Facebook was to implement source credibility prioritization. In this same scenario, if someone was to like or share an article, Facebook could first check its feedback for that particular site and/or author. If there is enough feedback, and the score is negative, Facebook could deprioritize the article on other people's news feeds by a factor of the magnitude of the negative score. Likewise, if the site or author has a positive score, Facebook could choose to display the article as it normally would, or maybe even prioritize it higher. From a utility perspective, this has the potential to significantly reduce negative utility brought about by fake sites. On the other hand, there does exist the potential to filter out some sites which many people may find to be perfectly acceptable, which contributes to false positives. In other words, this has the potential to significantly decrease negative utility, with a smaller side effect of decreasing some positive utility. From an overall perspective however, implementing source credibility prioritization has the net effect of decreasing negative utility.

In terms of the technical feasibility of this point, we that it is one of the lowest hanging fruits in the pursuit of diminishing the spread of fake news. Once possibility would involve Facebook storing a few metrics for a site based upon feedback received by readers. Facebook already has a way to flag particular articles as inappropriate. In the same way, Facebook could allow a reader to optionally rate the credibility of the article or the site in general, and use these ratings to prioritize and deprioritize authors and sites respectively.

Lastly, there is the idea of checking whether or not the author of an article is indeed a real person, and is indeed who they claim to be. We feel that this would be a much more difficult task for Facebook to implement. Additionally, while author authentication is a concept that would indeed help lower negative utility, the implementation of it would most likely cross the privacy line. This in turn increases negative utility. For example, Facebook may attempt to link the author of an article on some site to an individual account on Facebook for an added level of author verification. Since this has implications in terms of privacy, we feel that author authentication is not an angle that Facebook should approach this problem from.

2 Reading Beyond the Headline

When posting an external link to Facebook, information is drawn from the content inside `meta` elements of the page source by Facebook's web crawler and placed into a preview on the post itself. If an element is labeled according to the Open Graph protocol, the crawler will place it in the link preview on the site (**fbwebmaster**). With this system in place, every linked website will appear the same on a users news feed. Any site can determine how its posts show up when linked on Facebook, regardless of the quality of the content or site itself.

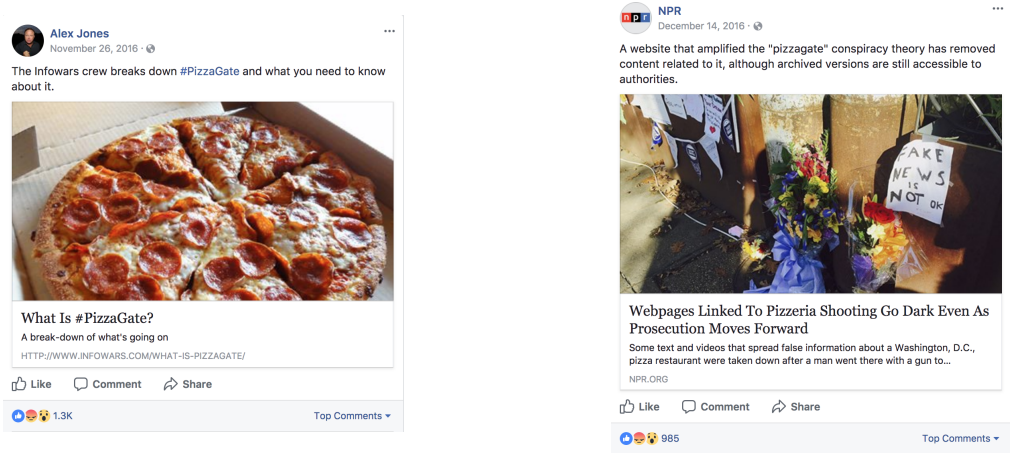

Figure 1: A comparison of posts by Alex Jones, creator of *Infowars*, (**alex˙jones˙pizzagate˙post**) and *NPR* (**npr˙pizzagate˙post**) about Pizzagate.

This uniform formatting treats *NPR* and *Infowars* or any other site that uses the `meta` elements according to Facebook's specifications as equals. It places less emphasis on the source of the article than on the headline and picture. The examples above are posted from official accounts, allowing users to easily see where the article comes from. However, the same articles posted by friends or unofficial pages make the source more difficult to determine. In addition, 50% of shared links generate less than 1% of clicks on social media (**clicks˙vs˙shares**), meaning most shares happen without the user clicking the link and reading the article. Therefore what a user sees on Facebook's preview is most likely all that user sees and is entirely determined by the web developer. With this in mind, there exists the possibility that the headline, picture, and description are totally unrelated to the actual content of the article, factually incorrect, or telling an attention-grabbing but incomplete story.

If Facebook were to put more focus on the content of the article rather than the title, this would reduce the negative utility generated by misleading or inaccurate articles being shared. In this case, automatic summarization is a potential solution, where an automatically generated summary is added to the preview, possibly in place of the description section created by the author. This

forces the text shown in the preview to be more representative of the actual content the user would read if they were to click on the article. An alternative would be to take the contents of the article itself and place a truncated version into the preview content on a user's feed. This would be easily understood by users with a short explanation, and potentially aid in users understanding the contents of the article before they share. In turn, this minimizes the negative utility generated by misleading headlines.

While Facebook could resolve this by implementing an automatic summarization of the contents of the page, summarizing text has many technical hurdles. Automated summarization of text is an active area of research in natural language processing (NLP) and even evaluating the quality of summaries in relation to longer pieces of text is often unclear. Summarizing often requires context for concepts, objects, or people mentioned. In addition, a system summarizing based off only the text presented to it can write summaries as misleading as what would have been in it's place without the automatic system. The truncated text could be subject to loopholes when gathering the text and potentially rife with abuse. Web developers could easily put misleading descriptions in text that is invisible on the page or only load text content from a server rather than being embedded in the page for Facebook's web crawler. With this said however, we still feel that implementing automatic summarization would lower the amount of negative utility.

3 Age of the Article

Situations like the Las Vegas shooting pose a difficult challenge for Facebook. In chaotic situations, important information can be gained from an individual's posts online, particularly before traditional media arrives. However, it is not clear until much later which posts are accurate and which are not. In instances where the posts are not true, included pictures often are of lesser known or regionally unknown celebrities. Fake posts of this type will increase negative utility for people trying to help locate missing people. However, there exists the possibility that limiting or preventing posts of this type might create more negative utility from angry or concerned friends and family. One possibility to combat both these issues could be limiting blog posts and unverified news sources from reaching the trending topics list for a period of time after the event occurs. This could be a way to minimize negative utility for users looking for information on the incident.

Facebook has recently implemented the ability for users to "check in" safe after a variety of events from around the world. The primary motivation for this is an attempt to minimize the flood of fake posts searching for friends and family without impeding anyone from actually making the posts. However, the effectiveness of this system is not known. Anyone who does not hear from someone they care about may still post in search of that person, leaving the door open for fake posts to flood a user's feed. Limiting content on the trending topics list immediately surrounding to verified accounts is very feasible for Facebook. With this said however, making a verified account

status more difficult to receive would be necessary, or maybe a different verification process for news stories or media companies would need to be developed. One potential would be to only allow trending topics from news sources with a high percentage of their content being previously verified. This could allow networks with a history of good behavior to get their stories to be seen without any interaction with developing stories.

While an enormous amount of negative utility is generated on Facebook every time an incident like the Las Vegas shooting happens, determining what is real and what is fake is not an easy process, even to potential human curators. The negative utility earned from implementing a bad classifier for filtering personal user's posts could easily generate far more negative utility than was initially created by the fake posts themselves. Fake or unverified news stories reaching trending lists might be easier to control if Facebook develops a method of curation that filters out unverifiable sources, such as message forums and blog posts.

4 Article Genre

Humorous content is popular on Facebook ranging from meme pages, to video skits, to satire news pages. The satirical media company, *The Onion*, has 6.7 million likes on Facebook, 700,000 more likes than *The Washington Post*. *The Onion* is a well known satire site and, upon first inspection, makes itself clear that it is always satire. However, *The Onion* as well as it's sister site *Clickhole* have a history of being mistaken for real news. In this case, the most well known online satire website has a notorious history for making headlines in national news. Furthermore, politicians such as Former White House Press Secretary Sean Spicer, took the site seriously (**spicey**). Other links from less well known sources are sure to cause even more confusion. There is some negative utility gained by users confusing satirical articles for real ones. Publications like *The New Yorker* can be more problematic because it is not solely a satire site, and simply publishes some satire, often without clear demarcation between what is satire and what is not. If something as labeled or appears like satire, it is not necessarily immune from being harmful. The conspiracies spread in the aftermath of the Las Vegas shooting were spread primarily by a message board on the website 4chan called /pol/, short for "Politically Incorrect". The board brands itself as satire and frequently excuses it's own bad behavior as a joke or satire.

If Facebook were to mark articles it detects as satire somewhere on the news feed, it may reduce negative utility for users who would be confused by the article while for users who enjoy the content, it would have minimal impact. Filtering satire altogether would create far more negative utility for users who enjoy the content than often brief spells of confusion caused upon initially seeing a satirical article. Allowing all satirical content onto the site, however, will also contribute negative utility.

A recent paper using satire to detect misleading news achieved 90% precision and 84% recall

with a relatively simple model (**satire˙detection**). This shows that, with further research, this could be a promising method of aiding users in determining whether news is misleading. However, determining whether an article is satire may not show the complete picture. Humor can still be malicious and simply classifying an article as satire may let some things that could be harassing or bigoted onto a user's news feed when it should have been caught by another filter.

Detecting satire is a rapidly improving area in NLP. With this said however, classifying text as satirical or not does not paint the entire picture and should not be a determining factor in whether an article is removed from Facebook or prevented from reaching a trending topics list. However, marking an article as satire might be a helpful tool if provided to users for when they consider which context to read the article in.

5 Interpretation Bias

Facebook recently came under fire for their manual curation of trending news topics when former Facebook employees were interviewed by Gizmodo about the trending news section. One curator said "There was no real standard for measuring what qualified as news and what didn't" and that they regularly avoided right wing sites such as *Breitbart* and *The Blaze*. When pressed for a comment, Facebook responded by claiming they "take allegations of bias very seriously" (**gizmodo˙fb˙news˙curation**). By framing bias as a negative, Facebook implies that they view themselves as a neutral entity and that they believe it is possible for their site and trending topics to be free of bias. Every stance on certain types of news generates negative utility for those who wish to see it but do not, and those who wish to not see it but do. For these users, the site is not working as optimally as it could. However, a neutral stance on fake news has other negative utility generated by the articles themselves. Many fake news articles have consequences that reach beyond just an individual user's news feed. The gunman attacking Comet Ping Pong in response to Pizzagate is an example of this. The conspiracy did not begin and end online. It included months of harassment, threats to employees and patrons of Comet Ping Pong, and ended with a shooting with multiple people who may not even use Facebook injured or dead.

There will always be negative utility generated by users who want to see the types of content Facebook disallows. With this said however, Facebook can minimize the negative utility generated by the types of content with consequences outside of Facebook itself. In order to do so, Facebook would need to decide what kinds of content are allowed on the site and be able to accept that any stance will be subject to criticism. For Facebook to consider what biases their employees and their platform bring to news curation, they first need to accept that they will always have bias. Bias is an unavoidable part of curating news. By choosing what to allow through content filters, even by banning things commonly banned in other forms of social media, such as pornography, we open the site to bias.

Facebook would have to first have an internal discussion among employees and shareholders of the company, to help narrow down what kind of platform they would like to have for users. The company itself would have to come up with a value statement and be able to stand by those beliefs in the face of backlash, or be open to finding a way to work with users and advertisers without compromising those values. This would be difficult, Facebook would likely need to prepare to protect employees such as community managers from harassment. With this said, examining what the company values as a whole is a crucial first step in deciding how to move forward on any other component of implementing a fake news moderation system.

6 Cross Referencing

The next method for assisting in classifying fake news on Facebook is to implement fact checking for an article. This involves using a web crawler to go around and potentially fact check the sites. More precisely, in an ideal scenario, when a user posts a link on Facebook, the page linked to should have been indexed, and there should be a metric associated with the page that conveys the trustworthiness of the content presented. If the score falls below a particular threshold, Facebook may deprioritize the link.

From a utilitarian perspective, we feel that on the surface, this method seems like it has the potential to greatly hinder the spread of fake news. If the content on a page is primary false, then deprioritizing it hinders it getting around the web. The major downside we see is that often, the articles linked to on Facebook are not formally written scholarly articles, but rather more casual ones. In these types of articles, opinions are presented as facts, and vice versa. This creates an ambiguity in how we define a fact. We feel that in this case, even if there existed an NLP oracle that could perfectly distinguish fact from opinion, the majority of articles would be flagged as false negative simply because of people's writing style. Overall, what little positive utility that this brings would be quickly overshadowed by negative utility.

In terms of the technical feasibility, the same argument can be made here. From an NLP perspective, it would be challenging enough to even write a method that could parse the text and pull out statements, let alone distinguish fact from opinion. Additionally, the context of the statement would need to go through sentiment analysis to determine the authors intentions for even writing the article in the first place. For this reason, we feel that Facebook attempting to distinguish fact from opinion is technically infeasible, and because of this, would result in an increase in negative utility.

7 Supporting sources

The final method for identifying fake news is to check supporting sources (if any) and make sure they abide by the same rules. In other words, this method attacks the problem with a recursive approach. When a user wants to share an article on the site, the fake news classification algorithm could take a depth first approach at calculating a trustworthiness metric. This can be done by first calculating the trustworthiest of the sources linked to in the article, and then using this metric in the overall trustworthiness calculation for the main article.

From a utilitarian perspective, where our main goal is to minimize the amount of negative utility brought about by fake articles, we feel that implementing this method has a few angles to look at the problem from. The core idea of this method is to first apply the other methods that we feel are appropriate. We next take a depth first approach, using these metrics to generate a metric for the primary article. We feel that for the most part, this would be an extra step in minimizing the spread of fake news. One point to keep in mind however is that by taking a recursive approach, we may be rating the credibility of an article based on the credibility of a article with multiple degrees of separation that has virtually nothing to do with the article in question. Also, as mentioned in *Credible Sites and Sources*, there doesn't exist a reliable algorithm to to flag a link with the context that it's being cited in. For example, if an author is citing a fake news article with the intention of pointing out the fake news there, sentiment analysis would need to be implemented so that the trustworthiness metric would somehow understand this context, and not use this against the article's score.

In terms of the technical feasibility of the implementation, there are a few issues that would make this point difficult to implement. The first hurdle is that there would have to be some kind of standardization to the method that sites use to track citations and references. Most sites that people read have a nonstandard citations section at the bottom of the page that consists of random bits of metadata and URLs. Although this conveys the required information to the *reader*, it fails to be parsable by a web crawler. Additionally, this depth first approach to calculating source credibility would involve storing a large amount of metadata.

8 Conclusion

With the increasing veracity of fake news across all of social media, particularly Facebook, and the consequences of letting misinformation spread rapidly and unchallenged, our utilitarian analysis has shown that Facebook has an ethical responsibility to come up with an effective and equitable strategy to combat its spread. Additionally, a Pew Research Center survey found 68% of Facebook users get news from Facebook, but only 33% of Facebook users get news from local TV or dedicated news sites and apps (**pew˙news**). Facebook earns more than twice the engagement in news

than any single non-social media platform, but Facebook does not hold itself to any standard for the integrity of their sources unlike most traditional formats. Because of its large impact on users, Facebook has an ethical obligation to attempt to stop the spreading of this fake news among its users.

As we've seen in our analysis of the methods of inhibiting fake news, there are many challenges involved in implementing these methods. Not only are some of these methods technically unfeasible, but implementing them in algorithm form may open up the doors to more negative utility than it actually stops. Additionally, with the volume of posts shared, it is not possible for Facebook to tackle these problems manually with human intervention. Algorithmic methods seem to be the most practical solution if human moderation and automatic moderation were equally accurate. Like we've seen however, there are just as many different struggles for developing an algorithmic fake news detection system. The primary issues that we see arising come down to privacy concerns, interpreting an author's motivations, and automatic fact checking. However, like we've shown here, we feel like some of these points would make a large difference towards the utility of Facebook as a whole, and believe that it is worth Facebook pursuing them further.

If Facebook were to decide to invest in an automated system for detecting fake news, the most important aspect in our opinion is transparency at every step of the process. In order to begin designing such a system, Facebook needs to determine what they value in a news source and be willing to stand by those beliefs in the face of adversity or openly change them in response to criticism. By filtering/deprioritizing sites and authors that promote racism, fake news, harassment campaigns and so on, Facebook is making a political statement that it will not contribute to the spread of sites that have a goal of hurting individuals. Although there are grey areas in the process of filtering news, the most important component in the matter is the transparency. As soon as Facebook chooses not to release its classification methods, it provides individuals with an excuse to point fingers at Facebook for bias.