Evidence of OverExposure in NLP showcasing the negative societal impact of fake news as seen from the “Comet Ping Pong” mishap

SAPTARSHI SENGUPTA

FALL 2018 semester, CS 5761 Take Home Final Exam

Word Count: 2,386 (including section headers), 2,366 (excluding section headers)

I. What is Exposure?

Exposure is defined as a kind of *confirmation bias* which is modulated according to the availability or lack thereof, of information or resources. In other words, it represents the effect which the scarcity or abundance of knowledge sources have on human beings be it with respect to general knowledge acquisition/confirmation or performing research. Exposure bias is mainly of two types viz. overexposure and underexposure.

Overexposure is the situation which arises from having an abundance of information and knowledge sources for a topic. This is most commonly seen in resource-rich languages like *English* [1]. What this class of bias does to researchers is it instills a feeling of safety when trying to test out ideas. Knowledge poor languages do not offer such comfort and as such, researchers tend to stay away from them. As a result of this bias, topics which are overexposed, benefit from the abundance of research initiatives and publications. An example of overexposure can be seen in the richness of a resource like the *English WordNet* [2] which has a total of 155,287 unique strings (nouns, adjectives, verbs, and adverbs). As mentioned before, overexposure tends to create a confirmation bias in people because of the abundance of information available for a particular category in a topic that they are interested in.

Underexposure is the opposite of the above condition. It can be defined as that type of bias which is created as a result of overexposure. As more and more research is performed on topics which have readily available resources, it creates a dearth of such in its surrounding areas because of the obvious difficulties associated with performing research in those directions. A perfect example of underexposure can be seen in the work being done by Dr. David Beck and Dr. Greg Kondrak [3]. They are attempting to study a language called *Upper Necaxa Totonac*. Beck claims that languages like these provide unique opportunities for studying the evolution of societies and neurological phenomena like how human beings think. Unfortunately, as very little to no work is done on these rare languages, it becomes a victim of underexposure and thus the technology required to process such are underdeveloped. A simple search for ‘english’ vs ‘bengali natural language processing’ on google scholar returns ~2,620,000 results as compared to ~14,700 for the latter, despite Bengali being spoken by about 261 million people worldwide [4]. This is another example of underexposure.

Exposure bias is a serious problem not only in natural language processing (NLP) but to all other branches of academia and beyond. With overexposure, researchers tend to gravitate towards the more glamorous or “*hot-topic*” areas of scientific inquiry because contributions in those areas would tend to garner more attraction and in turn raise their popularity. On the flip side, topics which suffer from underexposure, tend to get left behind. Thus, this creates a dearth of research in those areas. Should such a situation continue to persist, the divide between the two categories would continue to grow as a result of which bias in NLP technologies would become more pronounced. Ultimately, only a handful of publications would be available for underexposed topics which in turn would make researching them more difficult. This pattern will continue to persist unless underexposed topics start to see the light of intensive research.

II. So What Happened?

On December 4, 2016, a 28-year-old man named Edgar Maddison Welch, traveled from North Carolina to Washington to investigate a story he had read online [5-7]. The story talked about the presence of a child sex slave ring in a pizza restaurant called Comet Ping Pong as a part of 2016 presidential candidate Hilary Clinton’s nefarious agenda. Seeing himself as the savior of those children, Edgar decided to investigate the matter and thus found himself at the restaurant on December 4th. He carried a rifle with him inside the place and fired three shots. Luckily, no one was harmed as the bullets attacked only the interiors of the restaurant such as the walls and doors. After conducting his “investigation” and

feeling confident that there wasn't such a scene, Edgar surrendered to the police peacefully. Following a trial in 2017, he was sentenced to four years in prison.

In addition to this news affecting people like Edgar, it also took a toll on the owners of the restaurant. James Alefantis was Comet Ping Pongs owner. A few days before the presidential election was to take place, he received death threats and violent messages [8]. When he tried to understand why such a thing was happening to him, he searched online for the answers and much too his dismay, he found several articles badmouthing his restaurant. These weren't ordinary negative reviews about his restaurant's food. These were fake news articles created by different entities which posited that his establishment was being used to harbor child sex slaves and was connected to Hilary Clinton's presidential campaign [8].

Hilary Clinton in an address on Capitol Hill, came forward and condemned the proliferation of such news articles and claimed that more important things than politics were at stake i.e. the lives of innocents who unwittingly get caught in such collateral damage [9]. She talked about the difficulties of identifying such kinds of news but assured the public that researchers are trying their best to combat and eradicate this issue.

A major advantage of such articles, in general, is their ability to easily influence people's thoughts. As a result of this, it enables entities or publishers pushing out such news to monopolize certain demographics for their personal gains.

The reason I chose this particular issue as an example of showing exposure bias in NLP is because of the effect such kinds of news have on people. Even after confessing, Welch maintained that what he read was not fake news and still had a hunch that he might have been onto something [6]. Many others also agreed with Welch and dismissed the idea of this issue being treated as fake news, claiming that it was being done to cover up something bigger in play (conspiracy theories) [10]. This just goes to show how biased people can become when they read articles which agree with their beliefs. More often then not people manage to find such articles (fake) as a result of their overexposure or overrepresentation in the news and social media. Thus, this case makes for a perfect fit with the exposure bias paradigm because of how people tend to gravitate toward those things with which they can easily associate themselves with.

III. Why did this happen in the first place?

In order to understand why this problem occurred in the first place, we have to look back at the idea of overexposure and confirmation bias. Fake news articles are in abundance on the internet. Moreover, people generally tend to fall prey to such articles because they reinforce their beliefs [11]. This, in turn, provides them with more ammunition to argue for or against certain topics without ever finding out whether the articles they read had legitimate information or not. There are statistics which support this idea. From a post by *The Independent* [12], we find that from a study conducted on the delivery speed of news, fake news traveled six times faster than real news to a sample of 1500 people on Twitter. The same article stated that researchers at MIT claimed that almost 70 percent of the news which would be retweeted would be fake news.

The problem at hand requires novel algorithms for detection. However, to ask what NLP technologies "*cause*" such issues is more challenging. The reason is that I believe NLP is trying to solve the problem rather than create or influence it. This is because it is not in the best interest of NLP systems to "*create*" fake news. The entire motivation of NLP is to do good for the society. Thus, creating systems which produce fake news seems counterproductive. Furthermore, in my opinion, even in the most extreme of hypothetical scenarios, having a natural language generation (NLG) system create fake news seems like a lot of work with the resulting articles not even being strong enough to elicit a reaction from people. NLG systems are used for positive applications such as summarization of videos [13] among other things.

Human beings, on the other hand, are the ones responsible for this problem. They have much more incentive to create and propagate fake news. While monetary gains may seem attractive enough, fake news enables publishers to create social divides which in turn achieves far greater goals such as political.

The problem with human language is that it tends to be highly ambiguous. The current state-of-the-art systems only do so well when dealing with this ambiguity. It's tough for these systems to handle general text processing let alone fake news. This is where I feel our current NLP systems fell behind which in turn led to such a mishap. Fake news identification systems are still in their nascent stages. Thus, while it is disheartening to see such occurrences take place, it is understandable why they do so. The biggest giants in social media and IT such as Facebook [14] and Microsoft are putting more resources into developing algorithms which can quickly detect these articles and in turn prevent such mishaps.

The main reason why I think this problem occurred is because most fake news detection systems aren't geared towards real-time handling of such news. From [8], it was seen that "*#pizzagate*" (this was the handle being used to propagate the conspiracy stories surrounding places like Comet) occurred in about five Twitter posts per minute. If news such as this spreads at such an alarming rate, it was no wonder why people like Welch believed it because of the numerous articles claiming the news to be true. Furthermore, most fake news detections systems are still in their early stages and they rely on manually annotated data and try to build classifiers around them. This is not suitable for cases such as these where immediate action is required. Thus, I postulate that the issue occurred because these systems weren't capable of handling the posts in real time.

IV. So How do we address this?

Fake News detection has become a trending topic for research in modern times. So far, the state-of-the-art systems are able to detect fake news with accuracies between 85 and 92% [15]. While this might seem like a high enough accuracy, it still isn't high enough. This is because, for a task such as this, we want to be able to detect fake news as soon as it's published with almost near perfect accuracy. The main reason for wanting to do so is because of the poisonous nature of such news. Once these articles find their way into mainstream media, it is not long before everyone has heard about them and by that time, the damage has already been dealt. Thus, it is of the essence that an effective solution be synthesized especially for extreme cases like the above.

The solution which I propose is based on techniques used by researchers in the field to detect fake news and is two-fold. The first solution tries to detect fake news at the point of origin i.e. just before it is about to be published. The second solution is more general and tries to detect fake news from large samples of news data. While solutions to both categories are required, as can be seen from [8], stopping the rapid spread of fake news at the mouth seems to be more important.

Solution 1: Trying to detect fake news in real time would be quite a challenge because the system in use must be aware of recent events and can thus tune its predictive powers to address whatever is being written by the user. I propose that the hypothetical system have the ability to detect fake news on the basis of the *hashtags* (#) that a user includes in their posts. If hashtags are used (be it in real or fake news), the veracity of such posts can be verified with reasonable accuracy by performing a global search for such tags. The greater the number of authentic hits matching each hashtag in the post, the more confidence the system has on that article. Retrieving such hits shouldn't take too long because of the general high connection speeds ISP's tend to provide and advanced search algorithms employed in today's systems.

I also envisaged equipping online forums with knowledge about world events. Having such knowledge would enable them to fact check what a user has typed on the fly. The system could then

provide suggestions on that basis and warn users about their incorrect information dissemination or in extreme cases prevent them from posting.

Although these look like naive solutions, I feel that they could be used as starting points to detect such news in real time.

I fell that this solution could handle the problem which was described. This is because, when the article is about to be released, it would get flagged by the system as containing illegitimate information and would thus never reach mainstream attention. Furthermore, the articles which would attempt to spread such news would contain the controversial "*#pizzagate*" hashtag and would thus get removed. Thus, this solution appears to solve the problem.

Solution 2: For tackling the second problem category i.e. detecting which articles in a dataset are fake, I propose the following solution. The solution approaches the detection problem like any supervised machine learning classification task i.e.

- Determine a set of features to extract from the articles in the dataset. This is perhaps the most important part because the accuracy of prediction depends on the features being used. Features which could be used include n(1-3)grams, POS (part-of-speech) and non-textual features such as the number of hashtags, external links [16] etc.
- Train the classifier with the obtained feature vectors and obtain the accuracy of classification. Continue repeating these two steps until you get the highest possible accuracy.

The drawbacks to this solution are its reliance on annotated data and selection of appropriate features.

While this is not exactly a new solution, I would think that this would be suitable for this category of fake news problems where annotated data is already available.

REFERENCES

[1] Mohammad Nasiruddin. 2013. A State of the Art of Word Sense Induction: A Way Towards Word Sense Disambiguation for Under-Resourced Languages. arXiv:1310.1425. Retrieved from https://arxiv.org/abs/1310.1425

[2] Princeton University "About WordNet." WordNet. Princeton University. 2010.

[3] Research Profile: Computational Linguistics and Natural Language Processing | Faculty of Science. (n.d.). Retrieved from https://www.ualberta.ca/computing-science/research/research-areas/artificial-intelligence/research-profile-computational-linguistics-and-natural-language-processing

[4] The 10 most spoken languages in the world. Retrieved from https://www.fluentin3months.com/most-spoken-languages/

[5] The very real consequences of fake news stories and why your brain can't ignore them. Retrieved from https://www.pbs.org/newshour/science/real-consequences-fake-news-stories-brain-cant-ignore

[6] Pizzagate conspiracy theory. Retrieved from https://en.wikipedia.org/wiki/Pizzagate_conspiracy_theory#cite_note-NY_Times_2016-12-05-9

[7] In Washington Pizzeria Attack, Fake News Brought Real Guns. Retrieved from https://www.nytimes.com/2016/12/05/business/media/comet-ping-pong-pizza-shooting-fake-news-consequences.html

[8] Fake News Onslaught Targets Pizzeria as Nest of Child-Trafficking. Retrieved from https://www.nytimes.com/2016/11/21/technology/fact-check-this-pizzeria-is-not-a-child-trafficking-site.html

[9] 'Lives Are At Risk,' Hillary Clinton Warns Over Fake News, 'Pizzagate'. Retrieved from https://www.npr.org/2016/12/08/504881478/lives-are-at-risk-clinton-warns-over-fake-news-pizzagate

[10] Menegus, B. Pizzagaters Aren't Giving This Shit Up. Retrieved from https://gizmodo.com/pizzagaters-arent-giving-this-shit-up-1789692422

[11] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *SIGKDD Explor. Newsl.* 19, 1 (September 2017), 22-36. DOI: https://doi.org/10.1145/3137597.3137600

[12] Lies and fake news travel much further and faster than real news, according to study. Retrieved from https://www.independent.co.uk/news/science/fake-news-twitter-spreads-further-faster-real-stories-retweets-political-a8247491.html

[13] Using NLG to Auto-Summarize Digital Video | | Automated Insights. (n.d.). Retrieved from https://automatedinsights.com/blog/using-nlg-to-auto-summarize-digital-video/

[14] Romano, A. Mark Zuckerberg lays out Facebook's 3-pronged approach to fake news. Retrieved from https://www.vox.com/technology/2018/4/3/17188332/zuckerberg-kinds-of-fake-news-facebook-making-progress

[15] Yang Liu and Yi-Fang Wu. 2018. Early Detection of Fake News on Social Media Through Propagation Path Classification with Recurrent and Convolutional Networks. AAAI Conference on Artificial Intelligence. Available at: https://aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16826.

[16] Martin Potthast, JohannesKiesel, Kevin Reinartz, JanekBevendorff, and Benno Stein. 2017. A Stylometric Inquiry into Hyperpartisan and FakeNews. ArXiv:1702.05638. Retrieved from https://arxiv.org/abs/1702.05638