Big Data & Society: A Replication and Extension of "Engineering Political Misinformation"

Regina Catipon
MACS 40100
4/21/20

Code for this project can be found here: https://github.com/rkcatipon/macs-40100-project

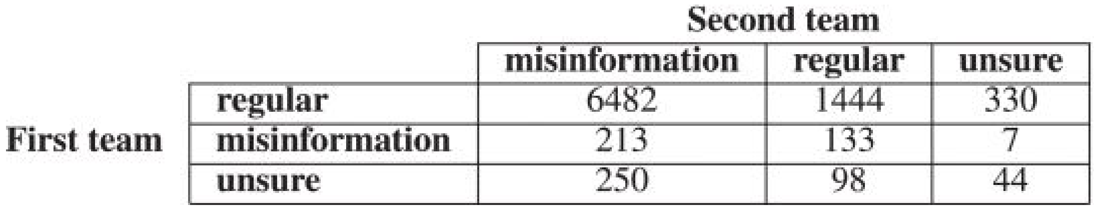
Introduction

In 2019, researchers from Imperial College London collected Twitter data related to the 2016 US presidential election over a span of four months after the election. Human annotators then labelled the tweets as fake news or not in addition to the type of fake news exhibited. The final publication, titled “Not All Lies Are Equal. A Study Into the Engineering of Political Misinformation in the 2016 US Presidential Election”, extracted the text and tweet features of misinformation, as well as the features of the accounts that share misinformation (Axel, Hua, Amador Díaz López,l Molina-Solana, & Gómez-Romero, 2019). The study demonstrates many admirable practices meant to create transparency and facilitate reproduction, such as the choice to make the dataset publicly available to all. Ethical concerns of the paper, however, include the possible introduction of bias with human annotators and whether researchers have an obligation to inform users and/or the platform of any misinformation detected in the course of the study. This paper will evaluate the study for signs of bias by replicating, testing, and extending the researchers’s quantitative methods. Then with these considerations, the ethical implications of reporting fake news tweets to Twitter will be assessed using a harm-reduction framework.

Quantitative Methods Assessment

Researchers have achieved remarkable accuracy in detecting fake news and misinformation. But no matter what method taken, from semantic to network-based approaches, training data often relies on expert annotations to set a model’s target labels and benchmark accuracy. As such, all such analysis depends highly on the concept of misinformation as understood by annotators and the subjectivity of their labels. Furthermore, fake news remains a relatively rare phenomenon with fake news making up only 6% of all online news consumption (Grinberg, 2016). Such observation sparsity may result in increased noise in a given dataset. Thus, even if a method has high accuracy, there are dangers of model overfitting and bias. When societies then adopt algorithms and big data analysis that contain such bias, these algorithms may contribute to the over policing of minorities and the targeting of the poor (Eubanks, 2018).

The authors of “Not All Lies are Equal” hypothesized that, “Deceivers strategically engineer their social media posts.” (Axel et al., 2019). In support of their hypothesis, they found that 1) fake news accounts engaged with others more than they produced content, 2) misinformation demonstrates different patterns in sentiment over time, and 3) that fake news tweets were more likely to be favorited. All of these results, of course, depend heavily on the quality of their label input. To address validity concerns around their annotations, the researchers had two teams label the tweets with the second team manually cross-checking every tweet with factual data (Axel et al., 2019). The two teams agreed on whether a tweet was fake news about a fourth of the time (fig. 1). A third team was then called upon to inspect the output of the second team and found it to be accurate. To “ensure the highest possible standard”, the authors only used the labels by the second team.



		Second team		
		misinformation	regular	unsure
First team	regular	6482	1444	330
	misinformation	213	133	7
	unsure	250	98	44

Fig. 1, "Agreement Between the Labelling Performed by the Two Teams in the Used Dataset", a table from the original paper.

The researchers demonstrate awareness of the drawbacks in human annotation by stating that they know their conclusions are limited by the subjectivity of the misinformation (Axel, 2019). Though they issue this caveat, the researchers could have actively engaged with and grappled with bias in order to understand how it affected their results. One way to engage with this issue of bias would be through an assessment of their quantitative methods. To test this theory, this paper replicates the sentiment analysis methods used in "Not All Lies are Equal" and then extends the text analysis methods with unsupervised topic models and network graphs. Text analysis methods like topic modelling can help to identify semantic patterns and detect latent structures (Blei, 2012) so through further text analysis, the researchers may have been able to find examples of bias in their tweet labelling.

Replication: Sentiment Analysis

The study leveraged both a classic lexical approach to sentiment analysis and a Deep Learning approach using *Word2vec* and *fastText* embeddings to train a classifier to predict negative or positive emotion (Axel, et al., 2019). To the authors credit, they encouraged replication and further analysis of their work, and chose a lexical approach that was easily reproducible with the R programming language, using the National Research Council Canada (NRC). Their traditional sentiment analysis approach found that, "misinformation tends to have a less positive sentiment" and that most of the calculated positiveness of the corpus came from the sentiment **trust** (Axel, et al., 2019). They found that **trust** was in over 40% of the fake news tweets.

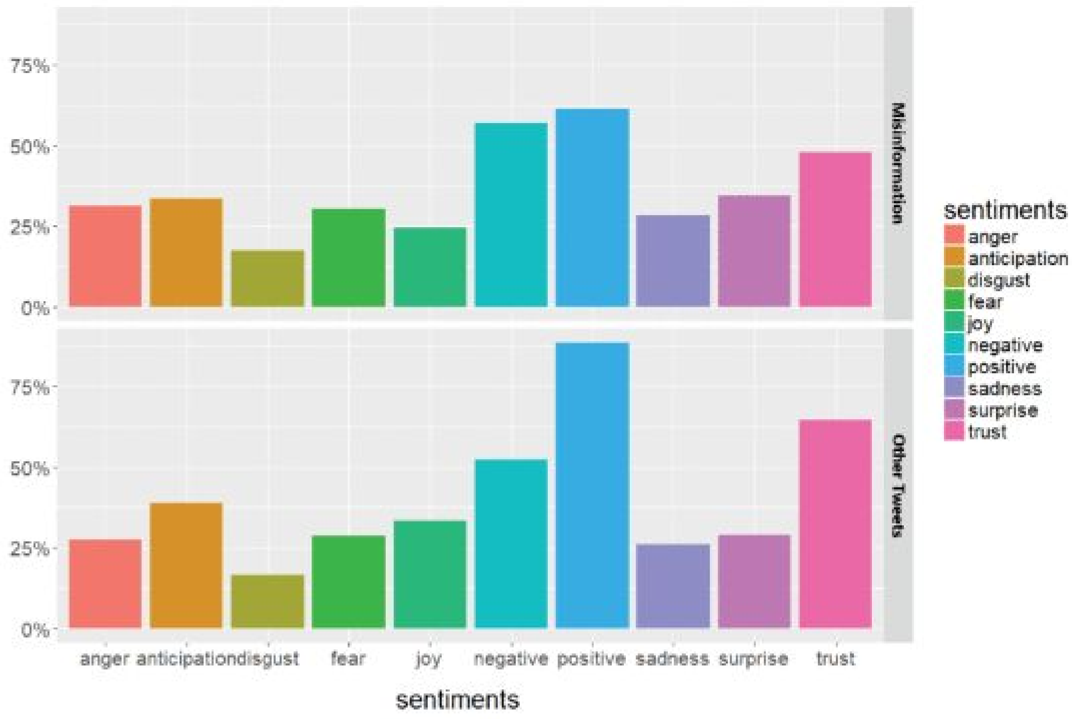


Fig. 2a, Original results for NRC lexical sentiment analysis.

To replicate their approach, each word was identified by sentiment and then the percentage of sentiment per tweet type was calculated.

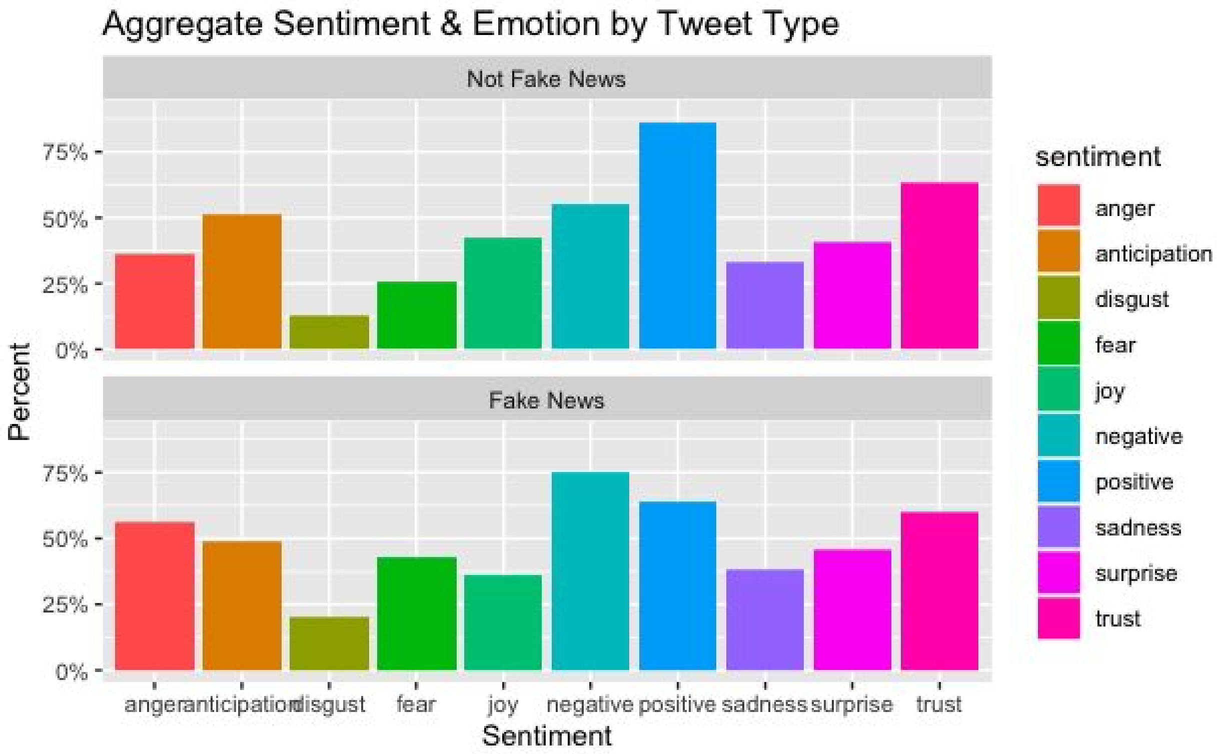


Fig. 2b, Replication results for NRC lexical sentiment analysis.

In replicating their methods, the fig. 2b above showed somewhat similar trends to those of the researchers and that misinformation did have less **positive** sentiment than its non-misinformation counterpart. The plot above also demonstrates that out of the NRC positive labelled emotions (not positive or negative labels), it was **trust** that had the highest percentage of tweets in the fake news subset. However, unlike the researchers, the replication results did not find an extreme difference between **trust** and other categories. Rather, **trust** at about 60% was closely matched by **anger**, 55% in the fake news tweets. The results do not provide any evidence of bias, as most of the trends are similar to what the researchers found.

And yet these results present no evidence of bias, it may be worthwhile to understand why the discrepancies between the replication and the original occur. Perhaps the authors used a different method for aggregation that would more strongly demonstrate that the two datasets of fake and other tweets are different. However, the table above also matches their results for novelty (for which they used the emotion of surprise as a proxy). The researchers state that they, "...found the mean sentiment of surprise for misinformation and other types of information to be 0.11 and 0.09 respectively." and the replication results also found this to be the same. So the discrepancy in sentiment counts may be attributed to perhaps a difference in filtering stop words or other coding approaches. Though sentiment analysis did not find cause to doubt the human annotation, it is possible that bias may be more visible with unsupervised methods such as Latent Dirichlet Allocation (LDA).

Extension: Unsupervised Topic Modeling

In order to keep the findings generalizable beyond the text and not limited to the use case of election tweets, the researchers focused more on distribution analysis such as looking at retweets over time. There are, however, still more they could have done with the text of the tweets such as generated Topic Models. LDA topic models have been useful tools for data exploration and discovery as they can help to detect latent structural and semantic patterns (Blei, 2012). To find the optimal number of topics, a perplexity score was generated on the full dataset. First, tokens with low term-frequency were filtered out, and then a document-term matrix was constructed to support probabilistic topic models. From there, perplexity scores were calculated to help determine the optimal number of topics in a model.

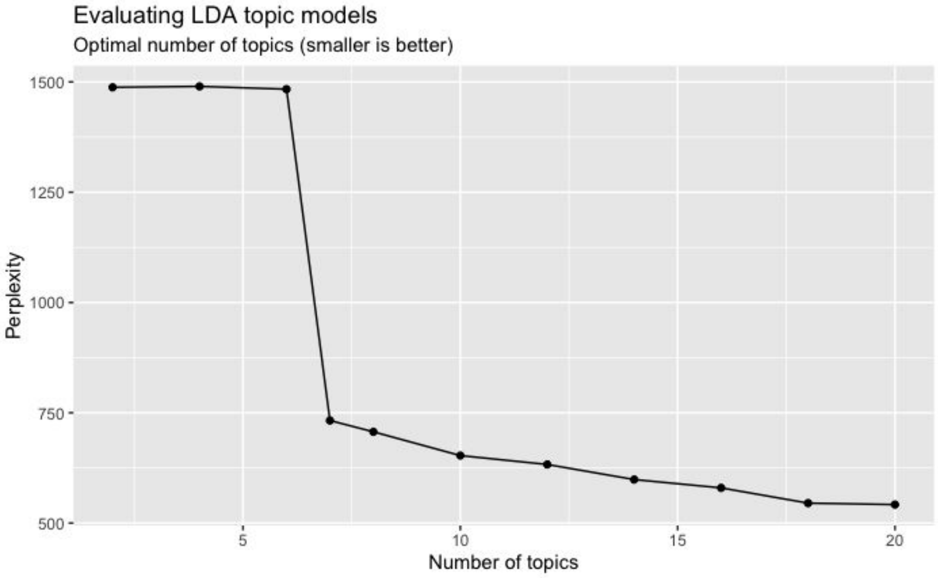


Fig. 3, Evaluating topic models using a perplexity score.

The LDA evaluation found a precipitous drop at topic six, where we begin to see slowing returns in perplexity as the number of topics increase until hitting a trough at 18. Because the number of observations in the dataset is so small, only 1,300 tweets, it would be best to choose a smaller number of topics or else risk having uninterpretable results. n = 7 was thus chosen.

Not Fake News Topic Model

The original dataset was subsetted into two, a fake news and other dataset. In the real tweets and misinformation subsets, low tf-odf terms at this time were not filtered out of a desire to capture rare terms as well. For the other (not fake news) dataset a LDA model was calculated with seven topics.

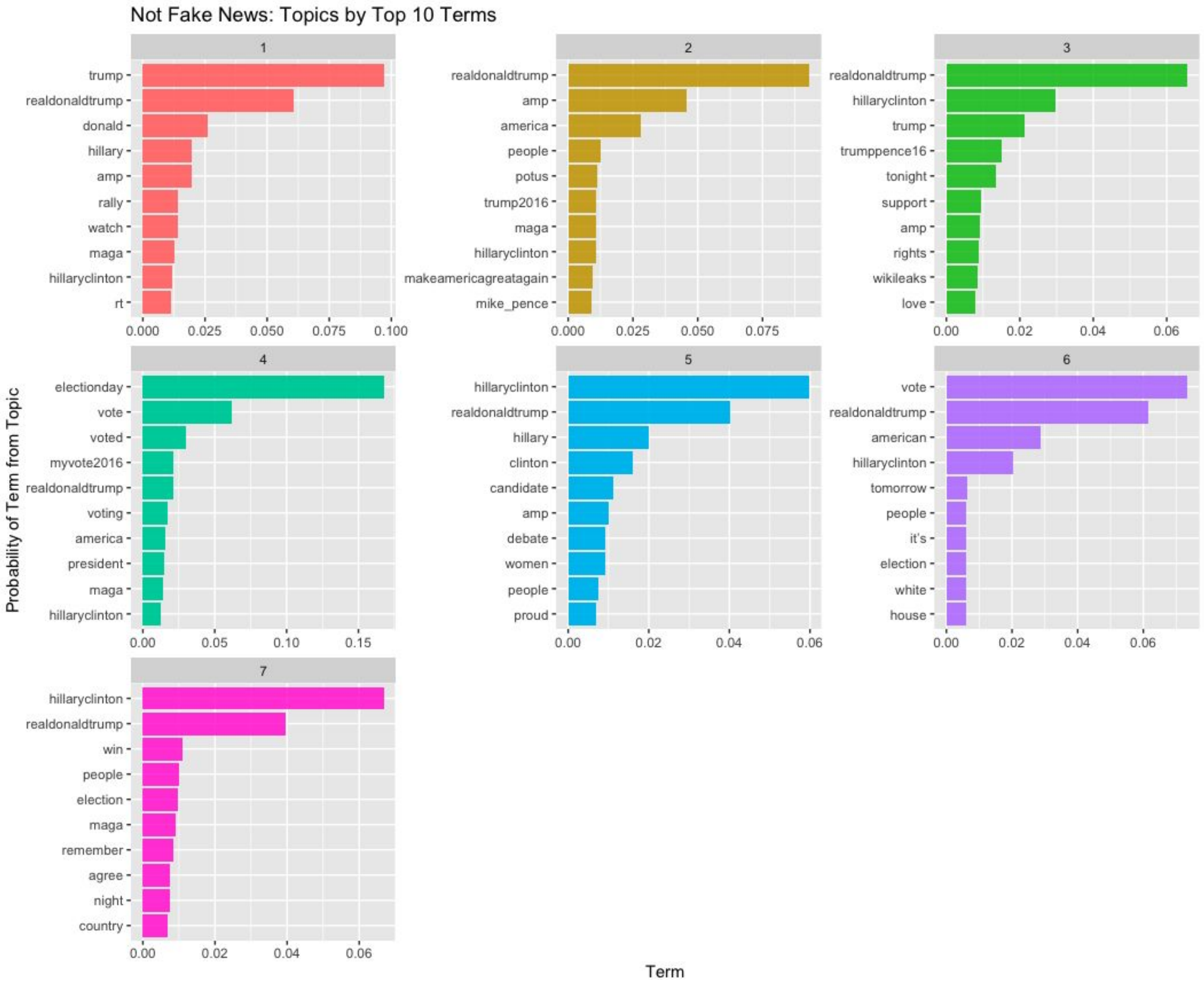


Fig.4a, Not fake news topic models show similar language with the exception of a few salient and defining terms.

The results show that terms such as *trump* and *hillary* overlap across topics. There are a few topics that seem to have salient terms such as **Topic 2** which seems focused more Trump focused with the terms *maga*, *makeamaericagreatagain*, and the hashtag for the campaign *trump2016*. **Topic 5** seems more focused on Clinton as a female candidate with *hillaryclinton* and variations of her name in the top 4 terms in the topic and the token *women.* Finally, **Topic 3** appears to be more focused on one of Clinton's campaign scandals from a right viewpoint with *wikileaks.* While there is some evidence of sub conversations in the online chatter, in total, the topics generated for the not fake news dataset seem to be semantically similar to one another. Now that we have some baseline of non-misinformation topics, this baseline can be used to compare the topics found in fake news.

Fake News Topic Model

To find features of the fake news dataset the topics generated in the not fake news dataset, a LDA model was generated and for internal consistency, seven was also chosen as the number of topics. First the text data was tokened and a document-term matrix was created for the fake news subset. Then the topic model was generated and the top 10 terms per topic were also plotted.

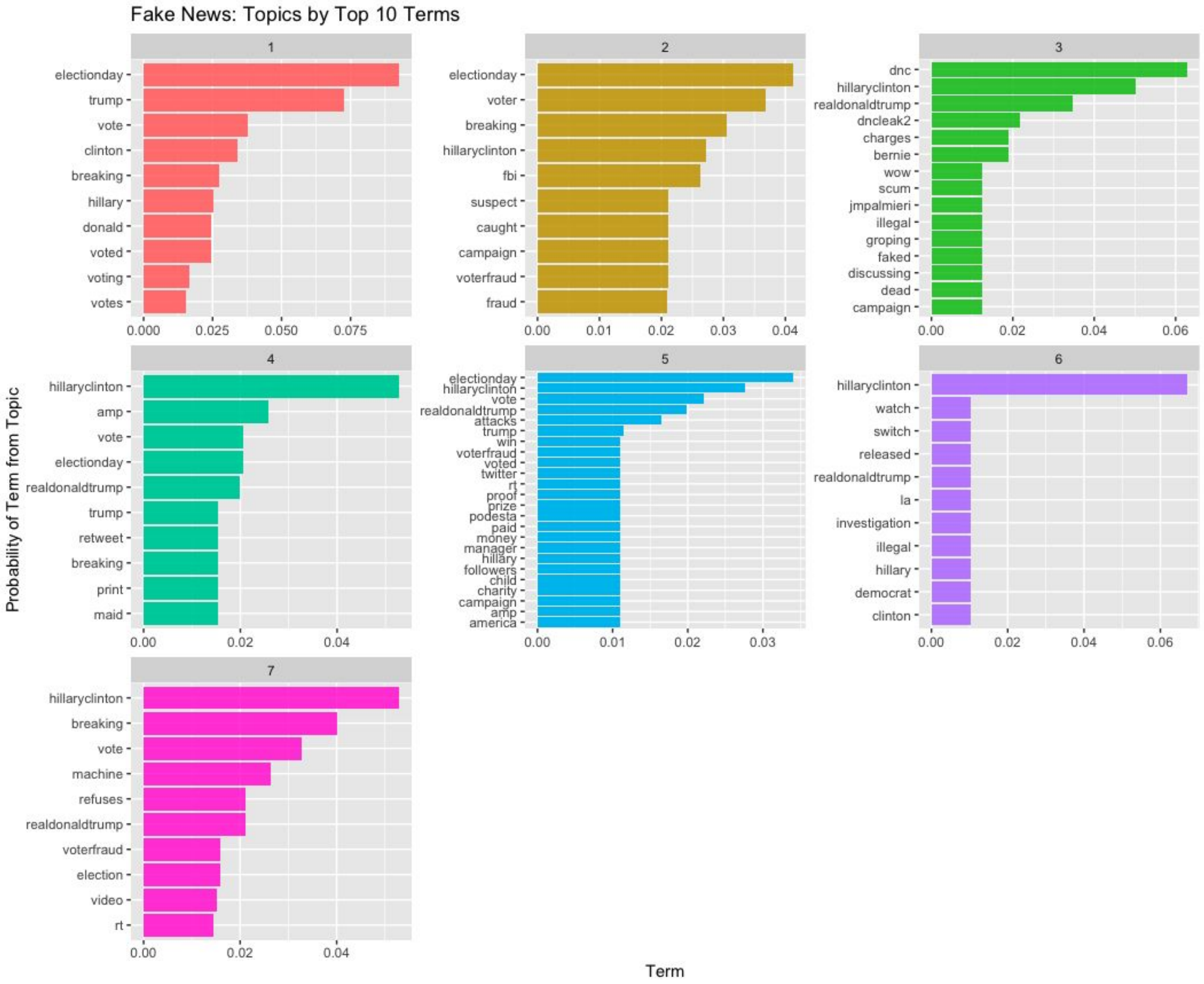


Fig.4b, Not fake news topic models show similar language with the exception of a few salient and defining terms.

Many of topics in the fake news model were focused on voter fraud or investigating Clinton. Of the topics, you'll notice that **Topic 5** has the most crowded y-axis. This is because many terms had the exact same estimated probability distribution loading. **Topic 5** - with the terms *proof*, *podesta*, and *child* looks to be focused on the Pizzagate conspiracy which believed that Clinton's campaign manager John Podesta ran a pedophilia ring that operated through a DC pizza shop. In **Topic 2** we can see terms like *suspect*, *fbi*, and *voterfraud* or *fraud*. Donald Trump is also absent from Topic 2's top 10 terms while *hillaryclinton* is 4th, suggesting that within this topic Clinton is more semantically linked to voter fraud. **Topic 7** also looks at voter fraud and points to concerns about the legitimacy of the voting process, a conspiracy also encouraged by now president Donald trump. According to the topics generated by the LDA model, we see that the language shown in the fake news tweets may be indicative of a news narrative that casts doubt on the US voting process or to raise more questions about Clinton's connection to illegal activity. These narratives are not as present in the other set of tweets, pointing to a difference in semantic styles between misinformation and not misinformation.

From the LDA topic model results, we see that it can be a helpful tool to identify differences between fake news tweets and the non-fake news tweets. Fake news tweets had more top terms that were related to conspiracies like Pizzagate and voter fraud. Interestingly, when comparing the two topic models, Clinton was more prominent in the fake news tweets, appearing as the top term in most of the fake news topics compared to the other dataset. If there was a target of the misinformation, it appeared to be Hillary Clinton and not Donald Trump. But whether misinformation that was anti-Clinton was more likely to be considered fake news by the annotaters is also a question of bias.

Not much is mentioned about the annotators, other than they were over 18 and did not receive compensation for the activity. Nor was much mentioned that of the third confirmatory group that adjudicated on the accuracy of the labels. It is possible that the annotations suffer from population sample bias. If there were college students, they would most likely be WEIRD -- white, educated, industrialized, rich, and democrat (Henrich, Heine, & Norenzayan, 2010). They may not, therefore, represent a diversity in viewpoints and may instead represent a homogenous view of misinformation. These annotators may be, for example, more likely to tag anti-Hillary Clinton content as fake news because it does not align with their world views. The point is, without further information about the annotators, it is difficult to know how the population was sampled and whether it was representative.

The authors are preemptive of criticism and do state that the paper is more of a starting point writing that, " It seems clear that different cultural backgrounds, knowledge of the American culture, and English language proficiency induced vastly different perceptions on whether a piece of information is considered misinformation or not." (Axel, et al., 2019). By running topic models, they may have seen this trend in anti-democrat candidate rhetoric in the fake news accounts. Could it be that misinformation steered more negatively in sentiment than the non-fake news tweets, because that is the nature of fake news? Or was the cause of the negativity due to the fact that most of the tweets were about women and perhaps women are treated more harshly on online platforms? Without examining more closely the discourse found in the text, their researchers do not address the ambiguity of what are the features of the text versus the features of misinformation.

Extension: Networks Graph

The paper also lacks clear evidence in its claim that misinformation was engineered and coordinated. The authors state novelty and polarisation ... "appear to be a product of coordination" (Axel, et al., 2019). It is true that others have covered coordinated spreads of fake news through networks (Vosoughi, Roy, & Aral, 2018), but it is not clear from the sentiment analysis results mentioned earlier that the higher levels of novelty are indicative of such efforts. One way the UCL researchers could have validated their expectations of coordination is by constructing a network graph.

Past social network graphs have shown communities of users that like and retweet one another, helping to boost the diffusion of fake news on Twitter (Grinberg, 2019).

Because the data provided did not include engagement information, we can not create a social graph network based on retweets or likes. It is possible, however, to create a mentions co-occurrence network where each node is an account mentioned within the text of a tweet, the origin point is the original tweeter, and the edge is the amount of times a user is mentioned. Such text-based networks graphs can demonstrate whether some communities contain handles that are affiliated with political groups. The political leanings of these communities may be useful in detecting bias in the tweet labelling. To explore the use of networks to detect annotator bias, a simple co-occurrence graph was calculated on the full set of tweets using Python's NetworkX package.

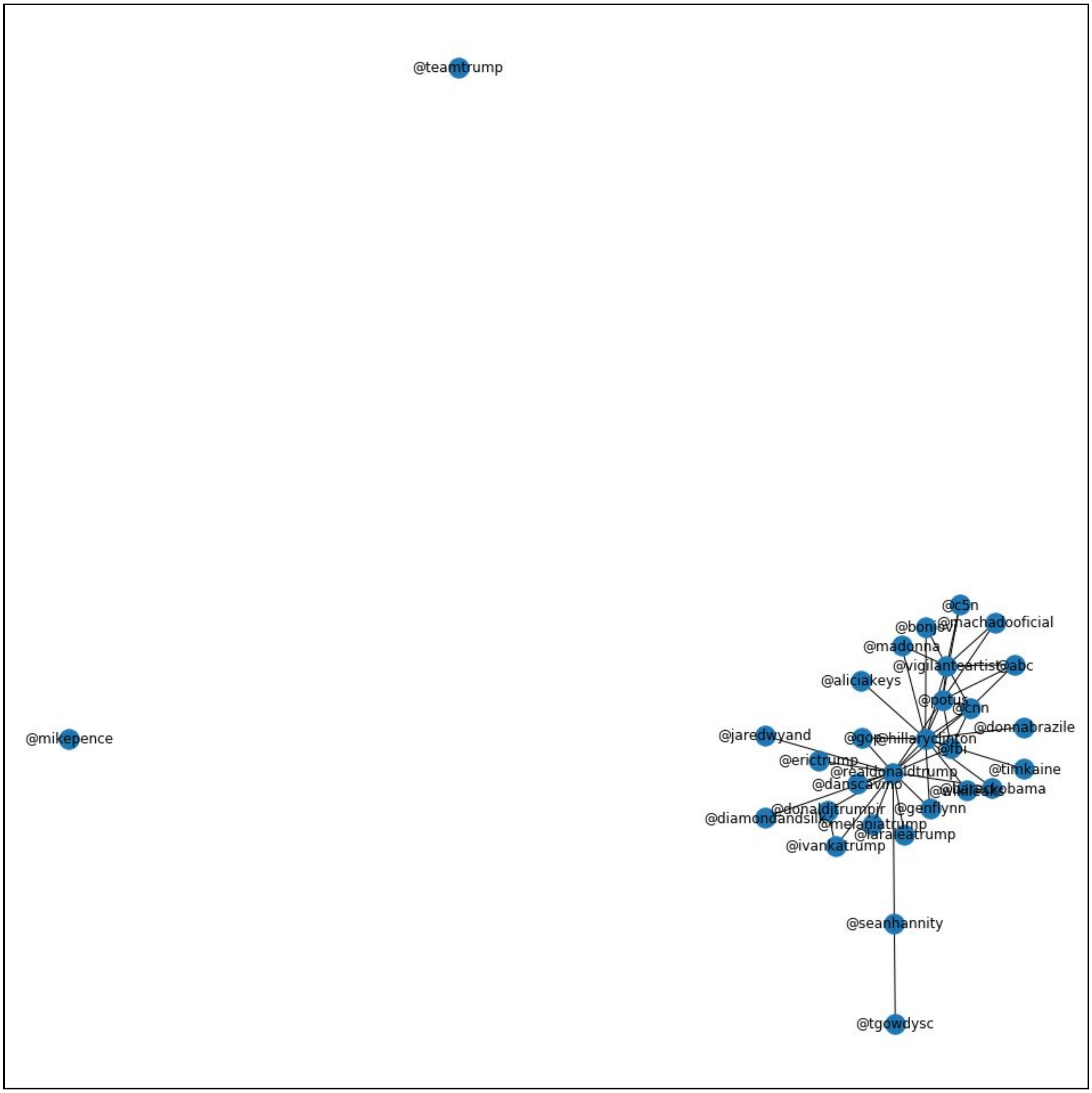


Fig.5, Co-occurrence NetworkX graph of mentions in tweets

From the results of the graph in Fig. 5, it seems that the structure of tweet mentions shows some clustering by political associations. Here we see that real @donaldtrump was mentioned in conjunction with @seanhannity and @ivankatrump. We also see that @hillaryclinton is also co-mentioned in tweets with @donaldtrump, but @donnabrazile and Alicia Keys are not co-mentioned with Trump. These node connections are aligned with the political leanings of the handles. At this point it would be helpful to construct a secondary networks graph on just fake news tweets to compare. It was not possible, however, to construct a graph based on a misinformation subset because the number of observations are so small at just over 100 tweets. And so, while there may be some potential in finding communities of misinformation and in validating the claims of coordinated diffusion, in terms of finding bias, these results are inconclusive.

Ethical Implications Assessment

In addition to questions of annotator bias, "Not All Lies are Equal" presents an ethical dilemma. When the dataset was released, should the users whose data was collected have been alerted that they have been exposed to misinformation? As of this paper submission, nothing suggests that users were informed by the research team or by Twitter despite the fact that screen names were shared. While interventions are uncommon in an academic study, a consequentialist and harm-reduction approach may justify notifying the platforms or the users.

An example of harm was actually identified with LDA topic modelling. In the topic model analysis, one sub conversation in the fake news tweets was found to focus on the conspiracy that implicated John Podesta with a child sex ring operating out of a pizza shop. The PizzaGate conspiracy theory eventually influenced a man to open fire on Comet Ping Pong in DC in December of 2016 (Lopez, 2016). If the users who shared that conspiracy on Twitter were notified that they were interacting with misinformation, then perhaps harm could have been reduced. It is, of course, easy to look back at these events with clarity of hindsight, but shooter scenarios have been one of the consequences of fake news on social media platforms.

Adopting a consequentialist view of the problem, it can be argued that to judge the morality of an action requires viewing it in regards to its outcomes (Quinn, 2017). Unlike deontology which states that morality lies in the action itself and not in its consequences (Quinn, 2017), according to consequentialism the effects of not alerting users in the PizzaGate example was immoral -- as inaction resulted in a mass shooting. Of course a consequentialist framework simplifies the event into a single cause and effect, and does not take into account the variety of other factors that contribute to mass shootings. But the framework also best illustrates why there should be urgency in notifying users of exposure to misinformation.

There are of course risks associated with alerting users that they have either interacted with or have been exposed to misinformation. One risk is that of false negatives, where misinformation is not identified. The other is the risk of false positives, where a tweet gets flagged or that a wrong account is implicated as spreading fake news. This may lead to an account being banned from the platform. To balance these risks, and given the fact that the researchers made the labelled dataset public, perhaps there should also be benefits of the study for the users in the dataset. In some ways, alerting users to their exposure satisfies the research ethic of Beneficence (Salganik, 2019) in that alerting helps those that are “harmed” by misinformation. To not notify users, however, means misinformation proliferates further on the platform and contributes to crises like mass shootings. If notification becomes an adopted practice, then it also raises the question as to whose role it is to alert the user and whose responsibility is it if the research misclassified fake news.

Recently, such roles have been clarified. As of May 2020, Twitter has decided to intervene and to mark accounts as potentially illustrating misinformation (Oremus, 2020). In order to reduce harm, researchers that detect misinformation should inform Twitter of their findings and leave it to the platform to adjudicate what requires intervention. By doing researchers may reduce the harm created by misinformation on platforms and beyond.

Conclusion

The initial questions around bias produced by human annotations were explored with additional text analysis methods. Rather than hard evidence of bias, the text analysis results suggest that the researchers need to disambiguate the characteristics of the dataset labels versus generalizations of the nature of misinformation. In regards to the ethical dilemma of notifying users, the assessment finds that because the dataset was made public thus increasing risks for the users, the researchers may have had more of an obligation to alert those in the dataset or the platform of the exposure to misinformation. This paper recognizes that to do so, requires the researchers to go above and beyond current practice.

Throughout the paper the authors emphasize how their findings should be considered “leads” rather than established truth. They make a concerted effort to be transparent in their methods, limit their claims, and share the data so others can replicate their work. From extending and replicating the methods of the original paper and then addressing ethical implications, this assessment concludes that while there is room for improvement, “Not All Lies are Equal” may in fact be a worthwhile lead in misinformation dissemination research.

References

Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press.

Lopez, G. (2016, Dec). “Pizzagate, the fake news conspiracy theory that led a gunman to DC’s Comet Ping Pong, explained”. Vox News. https://www.vox.com/policy-and-politics/2016/12/5/13842258/pizzagate-comet-ping-pong-fake-news

Grinberg, Nir, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. “Fake News on Twitter during the 2016 U.S. Presidential Election.” Science 363, no. 6425 (January 25, 2019): 374–78. https://doi.org/10.1126/science.aau2706.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world?. Behavioral and brain sciences, 33(2-3), 61-83.

Li, C., & Goldwasser, D. (2019, July). Encoding Social Information with Graph Convolutional Networks forPolitical Perspective Detection in News Media. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 2594-2604). https://www.aclweb.org/anthology/P19-1247.pdf

Oehmichen, Axel, Kevin Hua, Julio Amador Díaz López, Miguel Molina-Solana, Juan Gómez-Romero, and Yi-ke Guo. “Not All Lies Are Equal. A Study Into the Engineering of Political Misinformation in the 2016 US Presidential Election.” IEEE Access 7 (2019): 126305–14. https://doi.org/10.1109/ACCESS.2019.2938389.

Oremus, W. (2020, May). “Inside Twitter’s Decision to Fact-Check Trump’s Tweets” OneZero. https://onezero.medium.com/inside-twitters-decision-to-fact-check-a-trump-tweet-b5a30eaa3b1d

Quinn, M. J. (2017). Ethics for the information age. Pearson. (7th edition)

Salganik, M. (2019). Bit by bit: Social research in the digital age. Princeton University Press.

Schuster, Tal, Roei Schuster, Darsh J. Shah, and Regina Barzilay. “The Limitations of Stylometry for Detecting Machine-Generated Fake News.” ArXiv:1908.09805 [Cs], February 20, 2020. http://arxiv.org/abs/1908.09805.

Shu, Kai, Suhang Wang, and Huan Liu. “Beyond News Contents: The Role of Social Context for Fake News Detection.” In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, 312–320. WSDM ’19. Melbourne VIC, Australia: Association for Computing Machinery, 2019. https://doi.org/10.1145/3289600.3290994.