Online First
2021

Communication Design Quarterly

Published by the Association for Computing Machinery
Special Interest Group for Design of Communication
ISSN: 2166-1642

A machine learning algorithm for sorting online comments via topic modeling

Junzhe Zhu
University of Illinois
Urbana-Champaign
junzhez2@illinois.edu

Elizabeth Wickes
University of Illinois
Urbana-Champaign
wickes1@illinois.edu

John R. Gallagher
University of Illinois
Urbana-Champaign
johng@illinois.edu

Published Online April 26, 2021r
CDQ 10.1145/3453460.3453462
This article will be compiled into the quarterly publication and archived in the ACM Digital Library.

Communication Design Quarterly, Online First
https://sigdoc.acm.org/publication/

A machine learning algorithm for sorting online comments via topic modeling

Junzhe Zhu
University of Illinois
Urbana-Champaign
junzhez2@illinois.edu

Elizabeth Wickes
University of Illinois
Urbana-Champaign
wickes1@illinois.edu

John R. Gallagher
University of Illinois
Urbana-Champaign
johng@illinois.edu

ABSTRACT
This article uses a machine learning algorithm to demonstrate a proof-of-concept case for moderating and managing online comments as a form of content moderation, which is an emerging area of interest for technical and professional communication (TPC) researchers. The algorithm sorts comments by topical similarity to a reference comment/article rather than display comments by linear time and popularity. This approach has the practical benefit of enabling TPC researchers to reconceptualize content display systems in dynamic ways.

CCS Concepts
CCS → Computing methodologies → Machine learning

Keywords
Content moderation; Machine learning; Topic modeling; Forum design; Online comments

Manuscript received September 22, 2020; revised January 25, 2021; accepted February 15, 2021. Date of publication April 26, 2021.

CDQ 10.1145/3453460.3453462

INTRODUCTION
Ubiquitous technologies and scalable platforms have led to massive increases in content production. Daily, users send over 500 million tweets (Sayce, 2020) and upload 720,000 hours of YouTube video (Mohsin, 2021). In corporate platforms, user-generated content (UGC) can be produced far faster than it can be consumed or moderated (e.g., Amazon reviews, YouTube comments, Facebook posts, TikTok videos, and reddit threads). On a smaller scale, any website with a commenting functionality may suddenly incur large amounts of UGC. Moderating UGC in ways that benefit both back-end managers and users is thus an increasingly important issue and problem for technical and professional communication (TPC).

The most common solution to moderation at-scale has been to order and display UGC automatically through some sort of algorithm-driven model. While many algorithms are opaque due to proprietary or technical reasons (Burrell, 2016), we know that much UGC, such as online comments or reviews, is displayed by time, popularity, or some combination that uses both. For example, comments on YouTube videos are often sorted by either (a) displaying the most recent comment first or (b) by displaying the most popular ("liked") comment first. Likewise, Amazon.com reviews are sorted by achieving high scores of "helpfulness" votes, although even active writers of Amazon reviews are unsure of how these scores precisely function due to proprietary reasons (Gallagher, 2020). Displaying comments by time or popularity has been critiqued by media studies scholars for spreading disinformation and misinformation (Powers, 2017) and creating polarization (Shmargad & Klar, 2020). For example, Powers (2017) argues that users rush to comment on content first because that comment would likely be displayed first and/or at the top of all other comments, thereby making the comment extremely visible. More recently, Shmargad & Klar (2020) argue that algorithms use the "like" function, present in many digital networks, to order UGC; in their case, the UGC were comments made on political news stories. They found that comments automatically displayed via "like" functionalities polarizes users through reinforcement and disengagement (Shmargad & Klar, 2020, p. 424). Users' preexisting beliefs were reinforced by showing them comments that other users

in their networks "liked" or were disengaged because they did not see content outside their networks as few in their networks "liked" that content (Shmargad & Klar, 2020). Algorithmic alternatives to sorting UGC by time and popularity are thus necessary.

Using previous scholarship on technical documentation and content management in TPC, this article presents a preliminary topic modeling (Blei et al., 2003) approach for automatically finding relationships between online comments. Comments will function as our form of UGC for the remainder of this article. Using topic modeling, our model can group together comments that are similar in their content—not by time or popularity. While this model is only an initial entry into automatically sorting UGC, we believe it has practical potential benefits for both moderators and users. Moderators could use this approach to sort comments in ways that prioritize emerging real-time threads or find comments related to their editorial choices. Alternatively, if website programmers enabled users to sort comments related to ones users' favor, then users could sort comments in ways that were more useful to their preferences.

MODERATING USER-GENERATED CONTENT: EXIGENCIES & LITERATURE REVIEW

The need for moderating content on the social web has long been a concern for technical communicators (Gentle, 2012). Wikis, reddit, and other spaces driven by UGC must also have their documentation monitored (Gentle, 2012, p. 92–129). Monitoring threads and "conversations" (Gallagher, 2020; Gallagher, 2018; Gentle, 2012, p. 68) presents opportunities for technical communicators to take on the role of community managers (Frith, 2017). Extending Frith (2017), Cagle & Herndl (2019) advocate for managing online communities—specifically climate change forums on reddit—by gamifying forum engagement, nesting comments, and identifying the forum's ideal form of deliberation (pp. 35–36). Likewise, Pflugfelder (2017) identifies many exigencies that user questions may provide for technical communicators as managers and moderators. He examines "Explain Like I'm Five," which is a forum "…where plain writing and technical descriptions flourish, albeit outside of a professional setting" (p. 26).

While Cagle & Herndl (2019), Frith (2017), Gallagher (2018; 2020), and Pflugfelder (2017) focus on the more qualitative aspects of managing and moderating content for online communities, this article addresses the more quantitative and automated aspects of that managing and moderation. Specifically, we extend other machine learning attempts in TPC (Larson et al., 2016; Omizo and Hart-Davidson, 2016; Omizo et al., 2016) but with a focus on comment moderation for discussion forums. Using Amazon reviews publicly available from Stanford, researchers designed a prototype called "Use What You Choose" (Larson et al., 2016). This model used a support vector machine (SVM) that was guided by human coders to identify automatically rhetorical moves for instructional reviews (Larson et al., 2016, p. 6). Likewise, other researchers (Omizo et al., 2016) designed an application called the "Faciloscope" to help guide comment moderators for facilitating civil discussion by labeling comments as staging, evoking, and inviting. They used a set of human-labeled comments to train the Faciloscope.

While these approaches use human raters to train their machine learning models, our model more fully automates the identification process of comments by using latent Dirichlet allocation (LDA). Our model could be used to explicitly organize comments rather than identify rhetorical moves within the comments, which has been the main goal of previous machine learning approaches in TPC (Larson et al., 2016; Omizo and Hart-Davidson, 2016; Omizo et al., 2016). While the purpose of this algorithm is a proof of concept and preliminary, the algorithm sorts comments by topical similarity to a reference comment/article rather than display comments by linear time (comments made first are displayed first) and popularity (metrics such as "likes"). This approach has the practical benefit of enabling TPC researchers to reconceptualize content display systems in dynamic ways while attempting to inculcate a sense of trust in users by keeping the mechanism of automation transparent. This approach prompts TPC researchers to develop creative and innovative uses of algorithms.

METHOD AND PROCEDURES

Data collection of comments

The algorithm uses a large corpus of comments from the *New York Times* (https://github.com/JunzheJosephZhu/NYT-paper). In September 2015, the last author extracted comments from the NYT using its API function (https://developer.nytimes.com/apis) after requesting a developer key. All comments from May 1, 2015–August 31, 2015 were scraped; this means that a few comments were on articles from April 2015. This collection process resulted in 445,441 total comments (approximately 56% were initial comments and 44% were replies). We choose the NYT as a venue due to the newspaper's reputation as a "paper of record."

Topic modeling

After collecting the data, we ran topic modeling on the dataset using Gensim (https://github.com/rare-technologies/gensim/). We used latent Dirichlet allocation (LDA) to extract the topics (Blei et al., 2003), assigning fifty numbers to each comment, where each number represents the "weight" taken by a latent topic (which were then hand-labeled later in the study). Griffiths et al. (2007) write, "Representing words using topics has an intuitive correspondence to feature-based models of similarity. Words that receive high probability under the same topics will tend to be highly predictive of one another, just as stimuli that share many features will be highly similar" (p. 212). The LDA model is a technique used to find prevalent topics in a corpus of texts. The model assumes when each word from an analyzed comment is generated, a topic is chosen randomly according to a "topic weight" vector, and the word is selected from the chosen topic's distribution. Each comment has its own "topic weight" vector; each index indicates the probability that the corresponding topic is chosen when a word is generated. A topic, in turn, represents a distribution of words, indicating the probability that each word in the dictionary will be selected when the topic is chosen. The model is fitted to our corpus by iteratively updating the distribution of words associated with each topic based on the "topic weight" vectors for all comments in the corpus, then updating these "topic weight" vectors based on the newly inferred distribution. We manually labeled the theme of each topic (Appendix A) by looking at the words that are most likely to be selected given the topic.

In our experiment, we fit an LDA model with 50 topics. Therefore, for each comment in our corpus, we generate a 50-dimensional "topic weight" vector, consisting of 50 numbers that add up to one, indicating the probability that a word from the given document is sampled from each of the 50 topic distributions, while removing

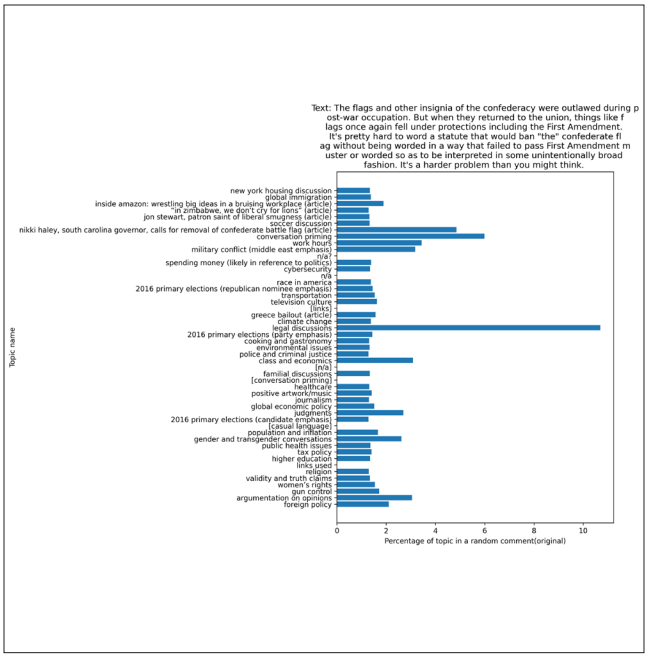

Figure 1: Topic vector for a random comment. "Legal Discussions" and "Conversation Priming" were the largest vectors.

other information, including length, sentiment, etc. In other words, each of the 50 vector entries represents the "topic weight", or salience, of a particular topic in relation to the comment. Figure 1 illustrates the topic composition of a random comment from our corpus. While a select few topics exhibit obvious saliency to a variable degree, most topic weights fall near a certain threshold indicating non-salient noise, which comes from the comment's non-significant association with unrelated topics. The most significant topic has a notably higher weight than all other topics.

We then grouped the comments by their most significant topic, which we retrieved by taking the maximum index for every "topic weight" vector. The number of comments under each group is shown in Figure 2.

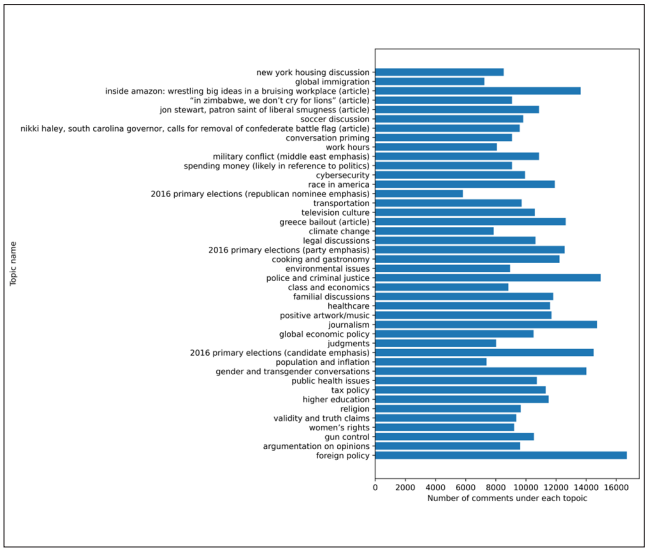

Figure 2: Topic vector for a random comment. "Foreign Policy" and "Police and Criminal Justice" were the largest vectors.

RESULTS: SORTING COMMENTS USING TOPICS

A practical application of this LDA algorithm is sorting comments using our topic modeling technique. Because each sentence can be represented with a 50-dimensional topic vector, we can find comments that are most relevant to the article itself/another comment (in the case of threads) by finding comments which have the smallest angular distance (computed using cosine similarity).

For a piece of reference text (article/comments at the top of a thread), we wish to sort a pool of comments replying to the reference text. Our algorithm first computes the LDA "topic weight" vector for the reference article/comment, and labels it *reference vector* (ref). Then the algorithm computes an LDA "topic weight" vector for each comment we wish to sort, which we label target vectors (target). For each target vector, we compute its cosine similarity with the reference vector, and label it the similarity score for the comment to be sorted. Formally,

$$score(ref, target) = \frac{LDA(ref) \cdot LDA(target)}{\| LDA(ref) \| \| LDA(target) \|}$$

where *ref* and *target* represent the text of the reference/sorted comment, respectively. Then the algorithm sorts the target comments by their similarity scores in decreasing order (Figure 3).

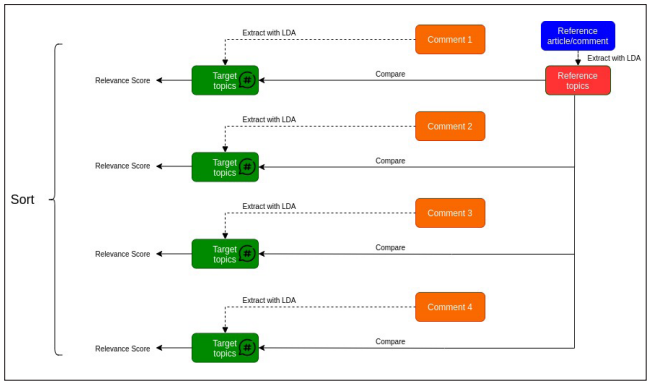

Figure 3: Flowchart of system that scores topics

Using this algorithm, content managers (Gorwa et al., 2020) and website designers can transparently gauge how relevant a comment reply is to a comment or even an article. Below are some examples of randomly selected comments, and their most similar comments gauged by topic similarity out of the entire dataset (Tables 1–4). All comments are verbatim except for the HTML tags that are the result of web-scraping. The tag
 is a tag for a line break.

The reference article talks about the labor conditions for white-collar workers in Amazon. It is a 6000-word article that covers diverse aspects of working at Amazon, including the pressurizing work environment, the data-driven philosophy, the requirement for dedication, and infighting among employees. As can be seen from the table, the top sorted comments address the topic of erratic work hours, valuable career skillsets, employee benefits, and corporate hiring strategies. The diverse topics of exhibited comments correspond to the diverse topics of the article itself.

Table 1: Five most relevant comments on the article, "Inside Amazon: Wrestling Big Ideas in a Bruising Workplace" from The New York Times corpus using the described mode

Comment	Cosine Similarity
Do you work 40 hours a week? Do you know when you'll be working? Are you well compensated for being on call or do you work for free when you tele-commute? Are you sent home without pay if it's slow or required to clock out but remain on the job site just in case there's an influx of customers? This article isn't about high tech industry or white-collar workers being on call. Everyone knows that there are certain industries that require this. It is about places like Wal-Mart, McDonalds, Home Depot, and other lower wage employers who schedule work hours so erratically that you don't know when you'll be working or even IF you'll be working any given week. Sorry, this article isn't about you!	0.956673029325405
Being an engineer is fine, but engineering or STEM fields relegate you to worker-bee status with some perks. I left my second job in engineering around 3 years ago at 26 and went into consulting with app/big data/good business skills and it has landed me a lot of fortune 20 gigs presenting my work to c-level/vp level at these companies. I had the same situation: I told my current employer, pay me 55% more or I'm leaving. The best they could do is 15%. I walked out the next day. And this is the job market for those with in-demand skills, just ask your price. Can you blame people in this position for abusing it? The thing is, most people in this position have honed their skills and put a lot of weekends of work in to get to the point which they are so much better than everyone else. I work 15–17 hour days, and while one might be off enjoying a Sunday soccer game with the kids, I'm most likely working. The idea here is, go get yourself some skills which allow you to have an extreme amount of power over employers.	0.95639873031407
I have a high tech job at a leading high tech F500 firm. Like most of our peer companies, our job performance is reviewed by our quality of work, not necessarily quantity. I rarely work nights/weekends unless I'm meeting with colleagues in opposite time zones to find a compromise. My stress level is very low, and my morale and motivation are excellent. I've had more time to spend with my family and personal life than any other time in my 15 year career because I've learned to manage myself and my manager. With all the talent and skillsets the fresh-out-of-college workforce can bring to a company, it's soft skills and personal skills (time management?) where they are severely lacking. My counterpart, who was hired one year before me, worked nights/weekends and was a constant suck-up to mgmt and a real go-getter whose plate was far fuller than mine was. This person was let go last year and those responsibilities were given to me. I haven't had to work a single after hour to wrap up those projects. A tip to the inexperienced: trust your team and respect everybody, don't be the smartest person in the room, delegate but share in success/failures, have your projects and deliverables under control and your boss will notice. Let your work speak for you.	0.955191178566295
I find it rather disingenuous that Bezos is not aware of this culture at his company. My gut tells me that he knows exactly what is going on, but that he just doesn't see that there is anything wrong with it. And those that are complaining are just whiners. Since more and more companies are going in this direction and many are expected to work at least 50 hours a week, I think its time for the Department of Labor to step in and provide some protection the way that non-exempt employees have. Too many companies are trying to skirt the law by making people "managers" or "contractors" so they can work them without having to provide compensation. And they need to take into account the technology tether that most employers now just expect. Already the wealth gap is exploding in favor of the 1% despite increases in productivity. But that isn't enough apparently. We need the DOL to come in and require employers to pay ALL employees and contract workers for the hours they work AND are required to be onsite, including time and a half and double time. These companies may become a better place to work if they have to think twice about demanding that extra report or that whatever at 4:00 pm on a Friday.	0.954151121942419
"1. In today's tech climate, some companies don't wish to pay their employees 150K or 120K for skills that can be outsourced for cost of 85-90K. Edison and Disney have outsourced some of their IT work, and quality of this outsourced work may or may not suffer." Next time you go to your bank and the ATM crashes or your supermarket and the checkout aisle is down because of a computer glitch you remember you said these words. The point is, as many of us can tell you, the quality is not there. And it's getting worse every day. For an example, we did an assessment for a large printer company that had hundreds of Indian contractors from Wipro. The company software had an invoice engine that took 24 hours to deliver an invoice. Imagine that. 24 hours to deliver a single one-page invoice. This had been developed by the folks from Wipro and by anecdotal accounts had worked like this for 2 years. We went in, and in 6 weeks of work (2 people) redid the software, added a few custom enhancements and the thing not only delivered an invoice in a few seconds, it had added functionality. We bid on another project and this large printer company decided to use Wipro. Does this make sense to you? If it does, you're a candidate for corporate management.	0.953911281263714

Example Comment 1:

"*The only way tax cuts are supportable, is if all of the business owners doing business with the government cut their prices, and their profits. Government contractors can't have their "growth", and their tax cuts too.*" (randomly selected)

Table 2: Comments most similar to Example Comment 1, a comment mostly about tax

Comments with similar topics	Similarity to original comment
You clearly have no clue. The amounts you are talking about were never earned in the US and no other industrial nation taxes such earnings. Any corporate manager who would earn money overseas and then bring it back to the US to be taxed should be fired, not praised. There is no "loss" to the Treasury except to those who believe everything belongs to the Government except which it deigns to allow you to keep.	0.96
The key is really in there, hidden in the verbiage. We should treat capital gains as income identical to wages. Then capital losses would count as a loss of income and be fully deductible on income taxes.	0.96
You fund private sector lifestyles with your spending choices just as much as you fund public sector employee wages with your taxes. And why do you care what people do with their wages, anyway?	0.96
This only helps if the employer would be required to pay the fee into a fund that would be entirely redistributed to any displaced workers. Having them pay the $50K to the Federal Government doesn't do those workers any good.	0.95

This example reference comment addresses the topic of tax policy. The top sorted comments are similarly mostly people debating about tax policies. The topics of sorted comments match closely with the topics of the reference comment, such earnings and government tax policy.

Example Comment 2:

"*I would have thought that stationing forces, even unarmed or lightly armed forces without organic transport so they wold be fixed in place and could not retreat, would have been the best tripwire to show U.S. commitment. But U.S. commitment without the support of allies may be too dangerous both to our forces and to politicians at home. Equipment can and has been abandoned and is not a very credible statement. Of course, if we want to create the belief that we are willing to risk military conflict after we have been shown to be unwilling and unable to impose credible economic sanctions, then we may be need to go back to something like the MAD doctrine that required leaving something to chance to be credible (we need to make adversaries believe we don't know what we are doing).*" (randomly selected)

Table 3: Comments most similar to Example Comment 2, a comment mostly about US foreign policy

Comments with similar topics	Similarity to original comment
Sorry Mr. McCoy, the Saudis are not "our partners in peace" either. Neither are the right-wing Israelis. Neither are our own neocons. We don't have any real "partners for peace" in the region right now. We've got to work with that we have, and that includes Iran as much as Netanyahu.	0.94
Israel has never used the US military or anyone else as a proxy. The issue is and always has been that those who have declared war on Israel and continue to openly threaten Israel are supported by the US, China, Russia and others. Therefore we, the US, play both sides of the fence. We support and arm Israel's enemies and arm Israel too. It is a total win for the US military industry.	0.94
Very, very well said p.kay. I agree with you completely. I believe that if this treaty is rejected Iran will rush to complete a nuclear weapon. I spent 1969 in Vietnam fighting in a senseless war that in 1968 Sec. of Defense McNamara declared as unwinnable. We must give this treaty a chance, we can always launch an attack. We may only have this one chance to avoid a cataclysmic conflict in the Middle East.	0.93
Israel, the only nation in the world unable to defend itself with force greater than that of those pledged to its destruction lobbing missiles at it's cities. It's the definition of antisemitism.	0.93

The reference comment addresses firstly foreign policy and secondly military conflicts. The sorted top comments address international relation issues and military issues to an observable extent. Therefore, the sorted comments are matched with the reference comment.

Example Comment 3:

"*I don't see why this is relevant. Is your point that since an ultrasound may be used to gather information for the doctor doing the abortion, that an extra ultrasound prescribed by the state legislature for propaganda purposes is not a problem? Do you think that if a mandated ultrasound is done that the woman's physician could simply use that one rather than arranging her own to pinpoint info she needs for the procedure?*" (randomly selected)

Table 4: Comments most similar to Example Comment 3, a comment mostly about abortion rights

Comments with similar topics	Similarity to original comment
Let's not forget that some of these pro-life folks were not above killing doctors and clinic workers. The campaign continues to harass and intimidate private providers of abortion services, by mailings, by websites, by sidewalk picketing. Many private doctors choose to work part time at Planned Parenthood, which can provide them the protection they need to take care of women.	0.93
When I was a kid and Roe v Wade happened, I asked my mom about it. She explained that rich and middle-class women could always get a (safe) abortion because they could afford it...meanwhile it was the poor women who suffered from back-alley butchers. Texas doesn't care about women's health. If it did it would streamline access for women's reproductive health (and prenatal care) across the board. Texas should just secede and does us all a favor.	0.92
By 'interrupt' you mean end (a pregnancy). Also, you're not the arbiter of how 'life changing' the decision to end a pregnancy is. For the vast majority of women who undergo the procedure at the hands of a competent medical practitioner, it is a tremendous relief.	0.91
If Planned Parenthood clinics are going to be eliminated, it is time for judges to rule that hospitals must perform them. It is a legal service and women are being denied health care and discriminated against. All those legislators who insist that abortion clinics must meet ambulatory surgical standards should be the first ones to co-sponsor the Hospital Abortion Availability bill.	0.91

The reference comment addresses the topic of abortions. The sorted comments also address the issue of abortions and pregnancy, in addition to mentioning Planned Parenthood. The chosen comments all seem to be pro-choice, while the reference comment does not have an obvious inclination. The algorithm sorts these comments together, which moderators and/or users could choose to group together.

DISCUSSION

As illustrated by the above examples, our technique is effective for finding relevance *between* comments. As a proof-of-concept, it therefore may be useful to moderators for automatically gauging the relevancy of replies in a thread, or the relevancy of a comment to an article. Additionally, using this technique, irrelevant comments/replies can be automatically hidden. Such a technique could also be an option for comment moderators to supply to users. Web developers could label this display according to platform needs.

There are three applications related to moderation for the algorithm we have proposed. First, the moderators could use this algorithm to find comments irrelevant to the subjects. This algorithm could thus assist moderators when they monitor comments threads. Second, the discussed algorithm presents an opportunity for users to read comments addressing a certain subject in an efficient fashion, thereby allowing users to *self-moderate* comments. Suppose a user is interested in a certain aspect of an article: with the proposed scheme, they would be able to select an arbitrary reference comment in which they are interested and then sort all other comments in terms of their relevancy to the reference comment (moderators could make this function available). The user would be able to read through all the comments most related to their interested aspect and output constructive discussion by replying to those comments. This interface could thus facilitate discussion on the aspect and provide positive feedback for user behavior. Third, when using the article itself as reference text, this scheme encourages users to keep the discussion closely related to the subject of the article itself and prevents diversion from the article content. By allowing users to selectively look at comments most related to the subject of the article, the algorithm may reduce distraction from irrelevant discussions. Moderation becomes essentially built into the scheme.

Researchers need not use the exact LDA algorithm we present in this article. Rather, using transparent automated techniques holds potential for unlocking diverse approaches to displaying online comments and their democratic potential. These approaches could help reorient participatory audiences that have become accustomed to opaque displays of proprietary algorithmic timelines and interface displays. Future researchers could use transparent algorithms to display comments and conduct qualitative studies about user experiences of these reading displays. To do so, other researchers might design websites that employ various algorithmic displays of comments. Researchers could then document user experiences via survey, interviews, and screen-recordings.

LIMITATIONS

Limitations of the data

The comments in our dataset lack threading and the model thus treats initial comments and replies as the same. Future researchers could develop a more integrative model that accounts for the interaction between initial comments and replies. For example, if threads are in the dataset, researchers can apply clustering on topic-specific sentiment analysis (Bhatia & P, 2018) to group comments into two sides of a debate. Generative adversarial networks (GANs) could also be developed with this approach.

Limitations of machine learning and machine reading

The topic analysis in this article is relatively simple due to our use of LDA modeling; we are simply using a bag-of-words representation. A topic modeling algorithm based on machine learning algorithms that takes orders of word sequences into account could be a more effective technique for topic modelling than our current model. Additionally, LDA gives each word a nonzero weight regarding each topic, as long as they appear in the corpus dictionary, no matter how trivial. This results in the 'noise' in the topic composition. A model that uses attention techniques (Vaswani et al., 2017) could produce less noise, thereby perhaps achieving more precise results.

Limitations with implementation

Our implementation of similarity metrics is relatively simplistic, using cosine distance as a similarity metric. A more accurate way could be treating the LDA vectors as embeddings, thus concatenating them and using a Siamese-network (Chopra et al., 2005) style classifier to measure their similarity. Despite these limitations, the algorithm we offer is transparent, as well as non-corporate and non-black boxed. It therefore has the potential to reorganize and redisplay comments in a manner different than linear temporality, popularity, or proprietary algorithm.

CONCLUSION: USING MACHINE LEARNING IN TPC RESEARCH

By developing algorithms for displaying comments in innovative ways, this study adds to research about content moderation, specifically comment moderation. Notably, our results show that comments can be coherently grouped together using topic modeling. Rather than rely on traditional methods of displaying comments, i.e., temporality or popularity, comments can be grouped through topic and cosine value. Displaying comments with this approach may help websites unlock the democratic potential of online comments.

More broadly, turning to transparent algorithms may help comment moderators to build automated systems in ways that create unexpected engagement by placing comments that are topically related but not necessarily temporally or popularly related. From this perspective, no human could arrange comments in this way—this approach is distinctly machinic. This model could be used to prevent users from manipulating comment organization, such as those who try to "gamify" comment displays or influence the perception of news by being the first users to make a comment. Such an approach to commenting could assist comment moderators with preventing misinformation or disinformation while keeping the moderating process transparent and open to the public. Transparent understandings of how and why automated systems organize information, in this case online comments, may yield inventive but productive opportunities for researchers of design and online interfaces.

ACKNOWLEDGEMENTS

The authors would like to thank Derek G. Ross and the two anonymous reviewers for their exceptional guidance and feedback, especially given the COVID19 pandemic.

REFERENCES

Bhatia, S., & P, D. (2018). Topic-specific sentiment analysis can help identify political ideology. *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 79–84. https://doi.org/10.18653/v1/w18-6212

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1), 1–12. https://doi.org/10.1177/2053951715622512

Cagle, L. E., & Herndl, C. (2019). Shades of denialism: Discovering possibilities for a more nuanced deliberation about climate change in online discussion forums. *Communication Design Quarterly, 7*(1), 22–39. https://doi.org/10.1145/3331558.3331561

Chopra, S., Hadsell, R., & LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR-05)*, 539–546. https://doi.org/10.1109/CVPR.2005.202

Frith, J. (2017). Forum design and the changing landscape of crowd-sourced help information. *Communication Design Quarterly, 4*(2), 12–22. https://doi.org/10.1145/3068698.3068700

Gallagher, J. R. (2018). Considering the comments: Theorizing online audiences as emergent processes. *Computers and Composition, 48*, 34–48. https://doi.org/10.1016/j.compcom.2018.03.002

Gallagher, J. R. (2020). *Update culture and the afterlife of digital writing*. Utah State University Press.

Gallagher, J. R., Chen, Y., Wagner, K., Wang, X., Zeng, J., & Kong, A. L. (2020). Peering at the internet abyss: Using big data audience analysis to understand online comments. *Technical Communication Quarterly, 29*(2), 155–173. https://doi.org/10.1080/10572252.2019.1634766

Gentle, A. (2012). *Conversation and community: The social web for documentation* (2nd edition. XML Press.

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society, 7*(1), 1–15. https://doi.org/10.1177/2053951719897945

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review, 114*(2), 211–244. https://doi.org/10.1037/0033-295X.114.2.211

Larson, B., Hart-Davidson, W., Walker, K. C., Walls, D. M., & Omizo, R. (2016). Use what you choose: Applying computational methods to genre studies in technical communication. *SIGDOC '16: Proceedings of the 34th ACM International Conference on the Design of Communication*. https://doi.org/10.1145/2987592.2987603

Mohsin M. (2021, January 25). *10 YouTube stats every marketer should know in 2021*. Oberlo. https://www.oberlo.com/blog/youtube-statistics#:~:text=500%20hours%20of%20video%20are,uploaded%20every%20day%20to%20YouTube.

Omizo, R., & Hart-Davidson, W. (2016). Finding genre signals in academic writing. *Journal of Writing Research, 7*(3), 485–509. https://doi.org/10.17239/jowr-2016.07.03.08

Omizo, R., Hart-Davidson, W., Nguyen, M.-T., Clark, I., McDuffie, K., & Ridolfo, J. (2016). You can read the comments again: The faciloscope app and automated rhetorical analysis. *DHCommons Journal*.

Pflugfelder, E. H. (2017). Reddit's "explain like I'm five": Technical descriptions in the wild. *Technical Communication Quarterly, 26*(1), 25–41. https://doi.org/10.1080/10572252.2

016.1257741

Powers, D. (2017). First! Cultural circulation in the age of recursivity. *New Media & Society, 19*(2), 165–180. https://doi.org/10.1177/1461444815600280

Sayce, D. (2020). The Number of tweets per day in 2020. *David Sayce*. https://www.dsayce.com/social-media/tweets-day/.

Shmargad, Y., & Klar, S. (2020). Sorting the news: How ranking by popularity polarizes our politics. *Political Communication, 37*(3), 423–446. https://doi.org/10.1080/10584609.2020.1713267

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł, & Polosukhin, I. (2017). *NIPS '17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6000–6010. https://dl.acm.org/doi/pdf/10.5555/3295222.3295349

ABOUT THE AUTHORS

Junzhe Zhu is an incoming Master's student at Stanford University Computer Science department. He received a Bachelor's degree in Electrical Engineering from University of Illinois at Urbana-Champaign. His current field of placement is audio and speech processing. He is interested in natural language processing and speech codec designs.

Elizabeth Wickes is a lecturer at School of Information Sciences, University of Illinois at Urbana-Champaign. Wickes' specialties include data curation, research data management, research programming, and the digital humanities.

John R. Gallagher is an assistant professor at the University of Illinois, Urbana Champaign. He studies interfaces, digital rhetoric, participatory audiences, and technical communication. He has been published in *Computers and Composition*, *enculturation*, *Rhetoric Review*, *Transformations*, *Technical Communication Quarterly*, and *Written Communication*. His monograph, *Update Culture and the Afterlife of Digital Writing*, is available from Utah State University Press. He co-edited, with Danielle Nicole DeVoss, *Explanation Points: Publishing in Rhetoric and Composition*, also available from Utah State University Press.

APPENDIX A: FIFTY MANUALLY LABELED TOPICS AND THEIR MOST RELEVANT WORDS

Topics (human-labeled)	Most Relevant Words
foreign policy	deal Iran Israel Russia agreement nuclear world war bomb Russian support country Jewish sanction Putin peace Iranian nation Israeli Cuba international Jew Ukraine west negotiate Palestinian attack leader power Japan
argumentation on opinions	question point answer fact wrong agree argument reason case base correct simply make suggest view assume completely clear argue simple statement disagree response opinion reasonable logic position side present give
gun control	gun people control kill fear act violence shoot stop hand carry fire murder death threat weapon dangerous happen attack NRA responsible public tragedy blame mass victim safety protect afraid safe
women's rights	choice life abortion choose woman support sex make marriage decision gay force pro baby issue decide reason marry person give anti matter couple relationship personal legal birth_control birth planned_ parenthood provide
validity and truth claims	fact lie claim truth true reality ignore real simply admit shame matter wrong prove public trust false face statement honest attempt hold call refuse excuse pretend hide accept action doubt
religion	religious religion freedom Christian church God belief society moral world culture faith practice Muslim evil pope modern Catholic respect base group accept true principle philosophy Islam sin follow concept form
higher education	school student college education high teacher public learn teach university class test degree year program kid graduate professor career math skill attend educate work major academic board job private grade
tax policy	pay tax government income benefit public fund cost cut program service federal raise money taxpayer social_security provide private low receive budget retirement revenue dollar earn spending fee welfare spend support
public health issues	drug exercise risk body healthy brain food weight disease pain eat effect study drink diet fat research vaccine calorie smoke problem sugar depression heart cancer normal cure alcohol condition physical
gender and transgender conversations	woman man male female young wear girl gender feel boy hair body sex dress sexual age face find ms fashion smile shoe clothe suit rape lady jenner transgender desire difference
population and inflation	high number increase rate low population large level average small result growth reduce compare percent rise standard due raise measure total percentage decline drop effect fall include factor inflation grow
2016 primary elections (candidate emphasis)	Trump candidate republican debate run GOP Hillary campaign Sander Bernie_Sander Donald_Trump win bernie presidential medium clinton fox primary supporter Biden appeal party chance nomination Hillary_Clinton politic speak establishment poll election
judgments	good make bad thing idea sense point great mistake hard give easy reason difference hope chance decision lot agree luck perfect pretty put happen sound part difficult mind place impossible
global economic policy	market China price economy business trade free economic Chinese profit sell buy corporation product industry interest corporate world stock big consumer create demand government investment feed financial global capital capitalism
journalism	read article time comment write story book nyt piece author reader report mention writer column paper New_York interesting news publish find post reading ms opinion newspaper reporter editorial journalist cover

positive artwork/music	love great hope feel life heart experience share bring find music hear wonderful enjoy mind listen eye moment art beautiful happy friend memory amazing learn sound remember song deep dream
healthcare	care health cost insurance patient medical doctor plan system provide treatment healthcare hospital aca cover Obamacare medicare service coverage physician good company therapist practice affordable treat premium therapy pay sick
familial discussions	child family parent kid friend life young age live mother father home adult grow son bear daughter wife husband generation raise single brother mom love sister die feel dad learn
class and economics	problem poor system rich solution people economic society social create policy class real middle_class poverty work blame solve wealth fix lack opportunity wealthy fail address continue inequality result elite part
police and criminal justice	police crime criminal case officer cop prison justice charge victim commit murder jail arrest judge abuse system stop sentence innocent violent shoot guilty behavior trial situation kill person evidence suspect
environmental issues	water land grow energy oil California clean produce resource environment power plant air source waste build farm farmer wind industry large big gas coal environmental area supply fuel natural sea
cooking and gastronomy	make food add eat good serve recipe easy restaurant cook bit great taste table tip cut fresh nice chicken meat store meal minute half dinner wine delicious hot egg top
2016 primary elections (party emphasis)	republican vote party democrat election voter democratic support political politician congress democracy GOP majority voting house interest elect win policy represent senate member representative senator issue politic poll base favor
legal discussions	law state rule court case legal constitution government decision act require Texas citizen pass federal justice lawyer protect judge congress apply robert decide intent process constitutional establish set enforce order
climate change	science study climate_change climate scientist datum change research show evidence base model rise earth scientific warm theory global temperature find record planet result effect expert paper trend year decade analysis
Greece bailout (article)	Greece Greek debt bank Germany Europe German economy country government loan European euro economic crisis financial money austerity pay krugman currency reform borrow banker default creditor leave union eurozone demand
television culture	show watch tv movie character funny video film picture miss story photo real laugh audience interview interesting scene joke episode season series review line star actor television news king reality
transportation	drive car line run travel walk road back stop mile train time foot front put horse fly driver park trip open check wait seat close ride hit jump place airline
2016 primary elections (republican nominee emphasis)	president Obama Bush run bill office state elect administration policy governor Clinton walker Reagan mr great fail jeb house remember presidency legacy promise Wisconsin christie Rubio Florida leader senator leadership
race in America	black white race American group people community African racism racist culture matter racial color identify person minority privilege america blow identity discrimination cultural society difference Hispanic individual skin Asian hate
cybersecurity	information security internet technology service phone secret email computer government datum personal find call ad access send public account record check apple provide online search site privacy private agency include
spending money (likely in reference to politics)	money make give big spend buy dollar time million put save lose sell worth back hand lot billion huge thousand amount waste hundred run politician real cash good small manage
military conflict (middle east emphasis)	war fight Iraq ISIS military force attack group enemy support middle_east terrorist world send government army destroy serve country Syria Turkey soldier Vietnam troop end conflict Muslim battle threat defeat

military conflict (middle east emphasis)	war fight Iraq ISIS military force attack group enemy support middle_east terrorist world send government army destroy serve country Syria Turkey soldier Vietnam troop end conflict Muslim battle threat defeat
work hours	day work time week hour hard leave back sit night wait put summer long stay office month room find end start sleep set full late break spend early turn walk
conversation priming	word speak hear comment find language refer mind guess sound describe call post mention puzzle explain fill clue point bit fit term note thought letter today kind phrase understand agree
Nikki Haley, South Carolina governor, calls for removal of confederate battle flag (article)	state history flag south slavery war civil today southern confederate_flag slave symbol fight represent fly remove American honor stand part nation hate display union north United_State red great racist proud
soccer discussion	play game team sport player great world good win fan football field big soccer match ball hit head final athlete watch top Fifa goal club run beat cup time professional
Jon Stewart, patron saint of liberal smugness (article)	conservative liberal medium political leave side wing view call anti hate social politic mr progressive brook idea extreme point ideology agenda left stewart center politician radical propaganda public daily voice
In Zimbabwe we don't cry for lions (article)	human life live world die death kill animal dog suffer nature lion end save survive people humanity planet hunt destroy specie loss dead cat understand alive protect natural human being cecil
inside amazon: wrestling big ideas in a bruising workplace (article)	job work worker company pay business employee wage union amazon hire labor corporate employer CEO Disney corporation management replace benefit salary tech customer profit low skill visa minimum wage executive employ
global immigration	American country America world nation citizen USA immigrant Europe United_States illegal immigration border Mexico foreign European India Canada national bear refugee Africa Mexican rest Fench bring France million Indian native
New York housing discussion	live city place move home area house build New_York local neighborhood building housing NYC town street community visit rent small resident neighbor mayor property project nice leave apartment large afford