

Countering Position Bias in Instructor Interventions in MOOC Discussion Forums

Muthu Kumar Chandrasekaran¹, Min-Yen Kan^{1,2}

¹Department of Computer Science, School of Computing, National University of Singapore

²Institute for Application of Learning Science and Educational Technology (ALSET)

National University of Singapore

{muthu.chandra, kanmy}@comp.nus.edu.sg

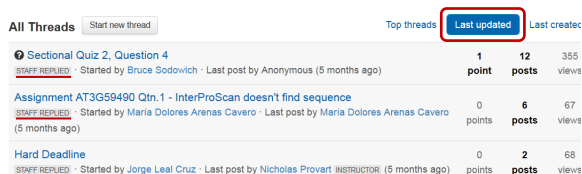
Abstract

We systematically confirm that instructors are strongly influenced by the user interface presentation of Massive Online Open Course (MOOC) discussion forums. In a large scale dataset, we conclusively show that instructor interventions exhibit strong position bias, as measured by the position where the thread appeared on the user interface at the time of intervention. We measure and remove this bias, enabling unbiased statistical modelling and evaluation. We show that our de-biased classifier improves predicting interventions over the state-of-the-art on courses with sufficient number of interventions by 8.2% in F_1 and 24.4% in recall on average.

1 Introduction

Massive Open Online Course (MOOC) platforms continue to evolve towards facilitating a better online learning experience. A key component of this effort is in platforms' ability to facilitate communication well, in part emulating the physical, face-to-face synchronous classroom experience. Despite debate on their effectiveness (Onah et al., 2014; Mak et al., 2010), MOOC discussion forums are still the primary communication medium for students to reach instructors.

In MOOCs, certain elements of traditional teaching are challenged by the scale of the class enabled by technology. The bandwidth of the MOOC instructor is especially strained given the high student-to-instructor ratio. Early research to address this gap proposed the problem of predicting instructor's intervention (Chaturvedi et al., 2014) in MOOC forums, as a means of aiding instructors in prioritizing their time towards productive intervention. That is, given historical account



The screenshot shows a forum interface with a table of threads. The threads are sorted by 'last updated' time, which is highlighted in a red box. The table has columns for 'points', 'posts', and 'views'. The first thread is 'Sectional Quiz 2, Question 4' with 1 point, 12 posts, and 355 views. The second thread is 'Assignment AT3G59490 Qtn.1 - InterProScan doesn't find sequence' with 0 points, 6 posts, and 67 views. The third thread is 'Hard Deadline' with 0 points, 2 posts, and 68 views.

Thread Title	points	posts	views
Sectional Quiz 2, Question 4	1	12	355
Assignment AT3G59490 Qtn.1 - InterProScan doesn't find sequence	0	6	67
Hard Deadline	0	2	68

Figure 1: Coursera's forum user interface used by both instructors and students lists threads sorted by "last updated time" by default. "top threads" and "last created" are other available sort options.

of discussion threads that were intervened by instructors, can a model learn to predict future interventions?

However, in this and follow-on studies on the same problem (Chandrasekaran et al., 2015b), there is a tacit assumption that what instructors actually intervene on is an optimal pattern of intervention. An underlying issue remains: Is there a difference between what instructors should intervene on and what they actually intervene on? Might there be systematic biases that influence the decision to intervene? While suspected, to date there has been no systematic study that proves that such bias exists.

Our study definitively shows that the answer is **yes**: instructors are biased and show suboptimality in their intervention patterns. Specifically, we show that instructor interventions in MOOC forums are influenced by position bias, akin to users of web search engines whose clicks on search results are biased by the order in which the results are presented (Joachims et al., 2005). Instructors view the list of threads being discussed on MOOC forums most often sorted by their "last update time" such as in Figure 1. We find that the distribution of instructor interventions over the positions of the sorted list of threads – the *positional*

rank – follows a log-normal distribution (see Figure 2). This implies that threads appearing at the top of the list are more likely to be intervened than those lower down. Given these defaults, observed ordering of items is time-dependent: the threads observed at one time can significantly differ between the different time points in which an instructor visits the forum. This effect, in turn, contributes to possible arbitrariness in an instructor’s decision to intervene.

The impact of this biased intervention is two-fold. First, the training and evaluation of statistical models that use the biased intervention data as in the previous work, is inaccurate. Second, the biased intervention decision may cause other intervention-worthy threads that appear further down the list to not be intervened at all. While previous work such as (Wise et al., 2012) propose alternative discussion forum designs to address the second problem the first problem deserves attention since large volumes of MOOC research data (e.g., the Stanford MOOC posts dataset (Agrawal and Paepcke, 2014))) has been collected from existing interfaces. In this paper, we propose methods to measure the bias and systematically remove its effects from a statistical model that learns the instructor’s intervention decision.

2 Preliminaries

Our corpus consists of discussion forum threads from 14 MOOC instances across different subject areas hosted by various universities across the world and taught by instructor teams of varying sizes on Coursera¹. In partnership with Coursera and in line with its Terms of Service, we obtained the data for use in our academic research². Table 1 shows our corpus’ demographics.

A discussion thread consists of posts by students, instructors, teaching assistants (TA) and community teaching assistants (CTA). Following prior work, we consider threads that are initiated by a student and replied to at least once by an instructor, TA or a CTA as an intervened thread. Threads started by an instructor are omitted since they are not interventions in a student discussion. Our problem is to predict interventions at a thread level, that is, the first post an instructor makes on a thread. So, we truncate intervened threads

¹Coursera is a commercial MOOC platform accessible at <https://www.coursera.org>

²However, we are unable to release the data for research without consent to release from the participating universities.

Course (-Iteration)	# of threads intervened	# of non- interventions	I. Ratio
BIOELEC	187	62	3.01
TRICITY-002			
BIOINFO	129	105	1.23
METHODS1-001			
CALC1-003	577	378	1.52
MATHTHINK-004	240	254	0.94
ML-005	883	1090	0.81
RPROG-003	359	738	0.49
SMAC-001	106	512	0.21
CASEBASED	25	96	0.26
BIOSTAT-002			
GAME	22	100	0.22
THEORY2-001			
MEDICAL	29	294	0.09
NEURO-002			
COMPILERS-004	15	601	0.02
MUSICPROD	2	228	0.01
UCTION-006			
COMPARCH-002	61	71	0.86
BIOSTATS-005	0	55	0.00
Total	2635	4584	

Table 1: Thread counts over the four main sub-forums of (*Errata*, *Exam*, *Lecture* and *Homework*) of each course iteration, with their intervention ratio (I. Ratio), defined as the ratio of # of intervened to non-intervened threads.

by removing posts after the first instructor post. We treat the problem of intervention prediction as a binary classification problem where intervened threads are positive samples and non-intervened threads are negative samples. We report the predictive performance of the classifier as F_1 score of the positive class.

We study threads gathered from Coursera sub-forums that are either self-identified or easily identifiable as contributing to the categories of *Technical Issues*, *Exam*, *Errata*, *Lecture* and *Homework* sub-forums. We omit others (e.g., *General*) as they are noisy with social discussions, or other reports on course logistics, irrelevant to the subject matter. To facilitate feature extraction we remove stopwords and replace occurrences of equations, URLs and video lecture timestamps with tokens <EQU>, <URL> and <TIMEREF>, respectively.

2.1 Baseline Classifier to Predict Interventions

We choose (Chandrasekaran et al., 2015b) as a state-of-the-art baseline system, hereafter referred to as EDM, for comparison. This system bettered the original (Chaturvedi et al., 2014) system in performance, and conducted work over a wider

Id	Thread Title	Last Update Time
971	In-video quizzes cannot be submitted	2014-03-24 20:46
968	Submit button does not work in one ...	2014-03-24 19:36
967	There is a typo or error	2014-03-24 19:35
966	When I click on Quiz submit button ...	2014-03-24 19:33
963	Duplicate lecture content ...	2014-03-24 19:15
957	Broken hyperlink in email	2014-03-24 19:12
902	Mistake in Q1 HW2	2014-03-23 18:17

Table 2: An intervened thread (ID 971) which was the last updated thread in this snapshot, taken at the time of its intervention. The forum user interface lists threads sorted by “last updated time” by default, introducing a position bias in instructor interventions. Note that thread with ID 962 is relegated to the bottom is perhaps a more important thread needing intervention.

range of MOOC instances.

EDM consists of a maximum entropy classifier, a type of linear classifier that handles feature spaces typical in text data, with several content-based features extracted from student posts in each thread. The features include unigrams (several thousands of features) from student posts weighted by its *tf.idf* score, the sub-forum type in which the thread appears, the length of the discussion thread in terms of number of posts, average length of posts, number of comments per post, discourse cues to the original post conveying affirmations, non-lexical references such as URLs to learner resources such as lecture materials, Wikipedia pages and timestamps in lecture video. The ratio of intervened (positive class) to non-intervened threads (negative class) is low (see Table 1). This class imbalance leads to poor prediction performance. To correct for this imbalance, they used class weights on examples, estimated as the ratio of intervened to non-intervened threads.

3 Measuring Position Bias in Interventions

To quantify the observed position bias on interventions we fit the data from a larger corpus of 61 different MOOCs, inclusive of the 14 MOOCs listed in Table 1 to different statistical distributions. For each intervened thread we obtain the snapshots of the list of threads ordered by their last modified time at the time of intervention. Using the snapshots, we count the number of interventions at each positional rank over all the courses, and fit this distribution of interventions over positional ranks against the power law and log-normal distri-

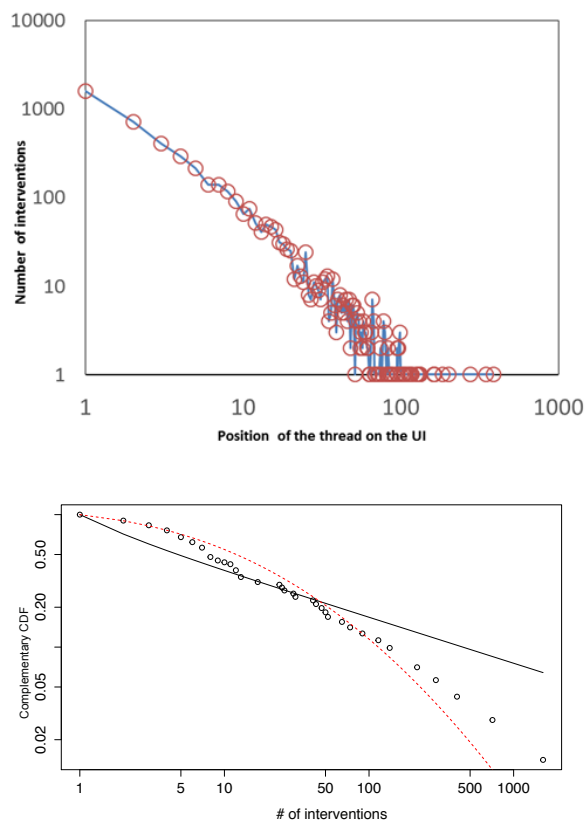


Figure 2: Log-log plots of (top) positional rank of threads vs. the # of interventions it received (bottom) the complementary cumulative distribution function (CCDFs) of the empirical distribution (circles) of interventions fit over a power law (grey line) and log-normal (red dashed line) distributions. Plots show interventions are clearly position biased and the log-normal (red dashed line) curve fits the distribution better.

butions.

We obtained the best fit for the log-normal distribution with parameters $\mu = 2.054$; $\sigma = 1.652$.
(0.196) (0.139)
Since our dataset is discrete we calculated the Kolmogorov-Smirnov (KS) goodness-of-fit statistic, $D = 0.143$, as prescribed by (D’Agostino and Stephens, 1986). Log-normal distributions are driven by multiplicative growth mechanism. It is typical in UI user log data where the attention (e.g., clicks) an object (e.g., search engine result) receives is proportional to the attention it already has. We did a model selection procedure to compare the goodness of fit of the log-normal distribution versus a power law distribution. We used the Likelihood ratio test (Clauset et al., 2009), where a positive sign on the log likelihood ratio with a $p < 0.1$ on the one-sided p -value rules out a bet-

ter fit to the competing distribution. Our results indicate that the log-normal distribution is a significantly better fit than a power law distribution ($-3.36; p < 0.001$; see Figure 2). The parameters of the distribution, μ and σ and the goodness-of-fit statistics, together quantify the position bias on interventions.

The above analysis shows that position is strongly correlated with intervention. This is not surprising; if instructors intervene often or if they can predict periods when intervention might be warranted (say, when an assignment is due), we should expect high correlation. To show that the position correlation leads to unwanted bias, we need to demonstrate that instructors intervene sub-optimally and favor intervening on results at the top at the cost of other, possibly more productive threads.

4 Does Position Bias Predict Intervention?

We ask if the signal from the position bias is strong enough to improve intervention prediction over the state-of-the-art (EDM). To test this hypothesis we model position bias as a simple, binary-valued feature set to 1 for a thread with a positional rank 1, and 0 otherwise. We augment this single feature to the feature set of EDM to create a new EDM+PB system. We compare the performance of EDM and EDM+PB individually over each of the 14 courses in Table 1. The models are trained on a random sample of 80% of the threads of a course and tested on the remaining 20%.

Table 3 shows the results from this experiment. On average, even this simple, position-augmented classifier improves EDM by a large margin of 13.7% in weighted macro average and 17.6% in simple macro average. CALC1-003 and BIOELECTRICITY-002 are notable exceptions where EDM+PB performs significantly worse than EDM. The intervention ratio of both these courses are above 1.0 (*cf* Table 1). We did not observe any decay in the numbers of interventions by position for these courses. Looking in depth, the instructors of these two courses may have monitored the forums continuously and tried to intervene on every thread, or may have also intervened without bias, based on the content.

The improvement on average and in the remaining courses is mainly due to increase in precision. This further indicates that the interventions are

Course	EDM			EDM+PB		
	P	R	F_1	P	R	F_1
BIOELEC TRICITY-002	76.9	60.6	67.8	100.0	24.2	39.0
CALC1-003	65.4	88.5	75.2	100.0	49.6	66.3
BIOINFOR METHODS1-001	35.3	26.1	30.0	100.0	56.5	72.2
MATHTHINK-004	36.8	17.1	23.3	100.0	48.8	65.6
ML-005	81.1	46.5	59.1	92.8	55.7	69.6
RPROG-003	47.2	50.0	48.6	67.3	51.5	58.3
SMAC-001	23.5	15.4	18.6	100.0	73.1	84.4
CASEBASED BIOSTAT-002	8.3	50.0	14.3	20.0	50.0	28.6
GAME THEORY2-001	25.0	14.3	18.2	100.0	57.1	72.7
MEDICAL NEURO-002	83.3	83.3	83.3	100.0	100.0	100.0
COMPILERS-004	33.3	50.0	40.0	33.3	50.0	40.0
COMPARCH-002	42.9	60.0	50.0	100.0	30.0	46.2
Macro Avg.	43.0	43.2	43.1	78.0	49.7	60.7

Table 3: Prediction performance of the position-augmented system EDM+PB showing significant improvement, over the baseline line EDM. Scores on musicproduction-006, biostats-005 are 0 due to low I. Ratio and are omitted.

strongly correlated with the position bias feature. Strikingly, on 8 out of the 14 courses, EDM+PB achieves a 100% precision. Examining the predictions in these courses, we found that the position bias feature was turned on in every correct intervention prediction, accounting for the improved performance.

5 De-biased Classifier

The EDM baseline does not account for the biased (non-) interventions. Due to the presence of position bias, thread instances thus vary in their *propensity* to be intervened. We need to counter the bias at the instance level. To implement this, we perform per-instance weighing with an appropriate classifier. We use SVM (Joachims, 1999)³, with the default linear kernel. We compute the per instance weights, w_{inst} , of intervened (positive) and non-intervened (negative) threads from two implicit signals respectively. They are (i) instructor’s propensity to intervene due to thread’s positional rank (ii) instructor’s confidence in discarding a thread from intervention.

Instance Weight Estimation. We estimate the propensity of a thread to be intervened from its observed positional rank. To discover an intervened thread’s positional rank at its intervention time t_i ,

³We use the SVM implementation SVM^{light} (<http://svmlight.joachims.org/>)

Course	Biased			De-biased		
	P	R	F_1	P	R	F_1
ML-005	56.7	61.6	59.1	55.1	79.0	64.9
RPROG-003	67.5	33.8	45.1	36.2	75.0	48.0
CALC1-003	63.5	93.8	75.7	61.5	92.0	73.8
MATHTHINK-004	39.3	26.8	31.9	47.4	65.9	55.1
BIOELEC TRICITY-002	77.8	63.6	70.0	77.5	94.0	84.9
BIOINFOR METHODS1-001	40	34.8	37.2	51.61	69.6	59.3
COMPARCH-002	43.8	70.0	53.9	44.4	80.0	57.1
Macro Avg.	55.5	54.9	55.2	53.4	79.3	63.4

Table 4: Model performance of the de-biased classifier Vs. a biased (SVM with class weights) classifier. I. Ratio on these courses are between 0.49 and 3.01. Best performance is bolded.

we reconstruct the snapshot of the thread (see Figure 2) listing at t_i . The number of interventions at each positional rank over all interventions was counted and normalised into probabilities. We then use the propensity of a thread to be intervened given its positional rank, $p(i = 1|r)$ to derive its weight, $w_{inst} = 1 - p(i = 1|r)$. That is, we weigh interventions that happen on threads with high positional ranks (i.e., towards the bottom of the user interface) as more significant and higher than those that occur on low positional ranks (i.e., towards the top of the user interface).

We also weigh non-intervened threads. We count the number of times a thread is skipped in favour of a different thread to intervene (# of snapshots where a non-intervened thread had appeared).

The resultant de-biased classifier (denoted EDM+DB) uses the same feature set used by the state-of-the-art-baseline, EDM. We compare its performance against a biased classifier, a system with the same feature set as EDM but without any instance weights. The biased classifier is equivalent to the EDM baseline.

6 Results and Discussion

The EDM+DB classifier varies in its performance in removing bias across different courses. To better understand its varied improvement, we examine its performance through three related questions.

1. *How well does the de-biased classifier perform?* Our de-biased classifier improves over the biased classifier on courses with sufficient number of interventions by 8.2% in F_1 and 24.4% in recall on average (see Table 4). We observe that the per-

formance of the de-biased classifier is sensitive to the number of interventions in the course. This is because the propensity score estimation (and the per-instance weights) are dependant on the number of times we can observe the state of the forum. De-biasing improves the F_1 on the high ratio courses in Table 4 (I. Ratio between 0.49 and 3.01), but does not improve F_1 performance for the 7 courses listed in Table 5, which all have low intervention ratios (less than 0.20).

2. *Can the de-biased classifier recover interventions that are missed by the biased classifier?* To be concrete, here we examine instances that were intervened by the instructor (positive), and identified correctly by our EDM+DB classifier (positive) but not by the biased classifier (negative). We randomly sampled 25 of 81 such instances that covered the courses in Table 3. The first author examined each of these threads and their instructor intervention using a taxonomy for interventions proposed by (Chandrasekaran et al., 2015a). This taxonomy grounded in pedagogy deems certain intervention types (e.g., justification request) are effectively made exclusively by instructors whereas certain other types (e.g., clarification) are optional for an instructor to make as peers can do them well enough. On this basis, the first author classified the 25 samples into those that warrant an instructor intervention and those that are optional. It was found that on 11 (44%) out of the 25 threads, instructor intervention was warranted. In the remaining 13, peers were actively answering the query, so we deemed these cases as optional for intervention. None of the threads were found to be solved or closed before the instructor intervened. We interpret this as a win for the EDM+DB classifier.

3. *Can the de-biased classifier identify thread instances that were not intervened due to the position bias?* Here, we examine instances that were not intervened by an instructor (negative), but were predicted to need intervention (positive) by EDM+DB. Again, we randomly sampled 25 of 42 such instances. As before we judged 9 (36%) instances as needing instructor intervention; i.e., we believe that instructors should have intervened, even though they did not. Two such instances are shown in Fig. 3. Another 8 (32%) instances had peer answers, which we deem as being optional. The remaining 8 were either solved or had social chatter that did not require instructor intervention;

Example 1: Thread Title: There is a mistake at 6:00 in the Week 3 Regularization - Cost Function lecture
Original Poster: The error can be seen and heard in the Week 3, Regularization, Cost Function lecture at the 6 min mark (image attached). The newly added regularization summation term in pink is written as the summation over variable i , but θ is subscripted with j . The summation should be over variable j . Andrew Ng also orally refers to “summation over i ” of that term, which again should be summation over j . The next slide shows a typeset version of the formula with the correct subscripts. Screenshot:
Example 2: Thread Title: PS6 #2
Original Poster: I missed this one so I thought I'd seek clarification. If a nonempty finite set X has n elements, then X has exactly 2^n distinct subsets. In the proof, the validation of $n = 1$ used the two subsets and itself. But I thought this was contrary to the statement “if a nonempty finite set”... Can someone help me understand this because set theory is definitely a weakness of mine. (various student answers follow ...)
Original Poster: I understand the empty set is a subset of every set, and I agree that the the Theorem is true, but in the proof, when element a is added to each subset, isn't it also added to the empty set, which would then create the situation of not having an empty set now? Just confused about how the proof handles the empty set situation with the ‘union U ’ procedure in the middle of the proof story.

Figure 3: Two threads that should have been intervened by instructors, where EDM+DB correctly identifies as needing intervention. Example 1 shows a thread that should be identified as an erratum report; Example 2 shows a thread where the original student poster expresses confusion that has not been clarified by any of the student answers.

but we note that such threads can easily be identified. Solved threads could be easily identified by attending to the last post made by the original poster, or the overall last post, both which typically provide the final answer to the original poster’s query and solves the thread. In one instance, the solved status of the thread was later indicated in the (updated) title of the thread, which could be easily captured.

We interpret this as major win for the de-biased classifier, as it can reliably pick out threads that have been overlooked by instructors that need intervention, with the false negative cases largely easy to correct using simple heuristics.

7 Related Work on Modelling Position Bias

Position bias due to the user interface and its effects on user behaviour has been observed in many domains. Much research on modelling or debiasing position exist, mainly in the context of web

search engines (Joachims et al., 2005; Pan et al., 2007; Craswell et al., 2008; Wang et al., 2016; Joachims et al., 2017) and recommendation systems (Schnabel et al., 2016; Liang et al., 2016). Joachims *et al.* (2005) and Pan *et al.* (2007) conducted eye-tracking experiments to confirm that users gaze at search results at top of the page and are, therefore, more likely to click them more often than the rest of the results. They observed that click behavior was biased, and does not always correlate with the relevance of the search result. Craswell *et al.* (2008) found that a cascade model – which posits that users examine results from top to bottom – best explained the bias. This examination pattern has been revisited by others (e.g., Liu *et al.* (2014)). In our study, we are interested in modeling user behavior and debiasing and correcting for it. Similar work has also been pioneered in the Web context. To model such observed user behaviour, improved ranking and click models were proposed. Joachims *et al.* (2005) proposed strategies to learn the relative preference between search results which are unbiased estimates of relevance. More recently, Schnabel *et al.* (2016) provided a generic framework to remove noise from biased training and evaluation data for recommender systems. Their algorithm learns disproportionately from items in recommendation systems according to their propensity to be clicked.

All the above works had access to reliable surrogate signals such as mouse cursor movements and clicks. In our MOOC scenario, we have only interventions (or lack of), recorded as instructor posts. Further, ranking frameworks assume a query to which retrieved items are listed in relevance order. In contrast, the default view of discussion forums are not ordered by relevance. While we cannot directly apply existing work to our setting, we draw from their inspiration and use the preferential judgement of the instructor to de-bias interventions.

8 Conclusion

We confirm the existence of *position bias* in instructor interventions in MOOC discussion forums and provide for a way to statistically quantify the bias. To enable accurate modelling and analysis we further proposed a de-biased classifier to counter for the bias and learn from biased instructor interventions. We show that the de-biased clas-

Course	Biased			De-biased		
	P	R	F_1	P	R	F_1
MEDICAL NEURO-002	100	83.33	90.9	66.7	33.3	44.4
SMAC-001	71.4	19.2	30.3	66.7	7.7	13.8
CASEBASED BIOSTAT-002	10.0	50.0	16.7	9.1	50.0	15.4
GAME THEORY2-001	50.0	14.3	22.2	16.7	14.3	15.4
Macro Avg.	33.1	23.9	27.7	22.7	15.0	18.1

Table 5: EDM+DB performance on low intervention (I. Ratio 0.0 to 0.2) courses compared against a SVM classifier with class weights, where each MOOC is evaluated individually. Best performance is bolded. Scores on compilers-004, musicproduction-006, biostats-005 are 0 due to low I.Ratio and are omitted.

sifier improves prediction when the training data consists of sufficient interventions. Importantly, the classifier can also identify clear cases where intervention is warranted but were overlooked by instructors.

We confirm earlier findings (Wise et al., 2012; Marbouti and Wise, 2016) on the bias induced by UI/UX. Since the effect of position bias, when extrapolated, can diminish students’ learning gains by compromising the instructor’s ability to judiciously intervene, we also call attention to the community to be mindful of the bias that UI/UX design can induce in MOOC platforms, intelligent tutoring systems and learning management systems, and to make design choices to mitigate this bias.

Acknowledgements

This research is funded in part by NUS Learning Innovation Fund – Technology grant #C-252-000-123-001, and by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office. We thank NUS Centre for Instructional Technology, Andreina Parisi-Amon from Coursera and Prof. Bernard Tan for helping us acquire legal permission to use Coursera’s data for our academic research.

References

Akshay Agrawal and Andreas Paepcke. 2014. The stanford moocposts data set.

Muthu Chandrasekaran, Kiruthika Ragupathi, Min-Yen

Kan, and Bernard Tan. 2015a. Towards feasible instructor intervention in mooc discussion forums. In *The Thirty Sixth International Conference on Information Systems (ICIS 2015), Fort Worth, TX, USA, Research-in-Progress*.

Muthu Kumar Chandrasekaran, Min-Yen Kan, Bernard CY Tan, and Kiruthika Ragupathi. 2015b. Learning instructor intervention from mooc forums: Early results and issues. In *Proc. of EDM*, pages 218–225. IEDM.

Snigdha Chaturvedi, Dan Goldwasser, and Hal Daumé III. 2014. Predicting instructor’s intervention in mooc forums. In *Proc. of the ACL (1)*, pages 1501–1511. ACL.

Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. 2009. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703.

Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An experimental comparison of click position-bias models. In *Proc. of WSDM*, pages 87–94. ACM.

Ralph B D’Agostino and Michael A Stephens. 1986. Goodness-of-fit techniques. *Statistics: Textbooks and Monographs, New York: Dekker, 1986, edited by D’Agostino, Ralph B.; Stephens, Michael A., 1.*

Thorsten Joachims. 1999. SvmLight: Support vector machine. *SVM-Light Support Vector Machine <http://svmlight.joachims.org/>, University of Dortmund, 19(4).*

Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2005. Accurately interpreting clickthrough data as implicit feedback. In *Proc. of SIGIR*, pages 154–161. ACM.

Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. 2017. Unbiased learning-to-rank with biased feedback. In *Proc. of WSDM*, pages 781–789.

Dawen Liang, Laurent Charlin, James McInerney, and David M Blei. 2016. Modeling user exposure in recommendation. In *Proceedings of the 25th International Conference on World Wide Web*, pages 951–961. Proc. of WWW.

Yiqun Liu, Chao Wang, Ke Zhou, Jianyun Nie, Min Zhang, and Shaoping Ma. 2014. From skimming to reading: A two-stage examination model for web search. In *Proc. of CIKM*, pages 849–858. ACM.

Sui Mak, Roy Williams, and Jenny Mackness. 2010. Blogs and forums as communication and learning tools in a mooc.

Farshid Marbouti and Alyssa Friend Wise. 2016. Starburst: a new graphical interface to support purposeful attention to others posts in online discussions. *Educational Technology Research and Development*, 64(1):87–113.

- Daniel FO Onah, Jane Sinclair, and Russell Boyatt. 2014. Exploring the use of MOOC discussion forums. In *Proc. of London International Conference on Education*, pages 1–4. LICE.
- Bing Pan, Helene Hembrooke, Thorsten Joachims, Lori Lorigo, Geri Gay, and Laura Granka. 2007. In google we trust: Users’ decisions on rank, position, and relevance. *Journal of Computer-Mediated Communication*, 12(3):801–823.
- Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as Treatments: Debiasing Learning and Evaluation. In *Proc. of ICML*.
- Xuanhui Wang, Michael Bendersky, Donald Metzler, and Marc Najork. 2016. [Learning to rank with selection bias in personal search](#). In *Proc. of SIGIR*, pages 115–124. ACM.
- Alyssa Friend Wise, Farshid Marbouti, Ying-Ting Hsiao, and Simone Hausknecht. 2012. A survey of factors contributing to learners’ listening behaviors in asynchronous online discussions. *Journal of Educational Computing Research*, 47(4):461–480.