The Impact of Textual Online Harassment on the Performance of Projects in Crowdfunding

Research-in-Progress

Wei HU **J Leon ZHAO**

Abstract

In the consequence-free and anonymous online environment, online harassment has become a serious problem. In many crowdfunding platforms, there exists offensive speech on the project pages, which might force potential funders to leave the discussion and to give up investment. The effect of online harassment on project performance remains unknown. This study attempts to investigate to what extent the textual online harassment score and the project creator's attitude towards textual online harassment might affect project performance. We constructed a Kickstarter panel dataset consisting of 388,100 projects and designed a novel framework and an algorithm BiLSTM-CNN to extract the textual online harassment score from comments, which can reach column-wise mean ROC AUC of 0.9463. This study contributes to crowdfunding and online harassment literature and provides important implications for reputation management of projects and crowdfunding platform design.

Keywords: Crowdfunding, online harassment, machine learning, bidirectional LSTM, Kickstarter

Introduction

Crowdfunding is the practice of funding different types of projects (for-profit, cultural, or social projects) by raising a small amount of money from individual founders without standard financial intermediaries, especially via the Internet. Nowadays, for lots of entrepreneurs, crowdfunding has become an alternative to traditional venture capital investments. (Mollick 2014; Schwienbacher and Larralde 2010)

Due to the no-intermediary property, the project creators need to promote and communicate directly with project backers. Sometimes, there will be disputes between the project creators and backers. Some aggressive and insulting comments are left on the project pages, which might leave uncomfortable impressions on visitors of the project pages. Project backers are afraid of fraud projects. They would doubt if there are integrity issues of the project pitches and project owners. Meanwhile, project creators are annoyed with unreasonable and malicious comments as well. Harassed comments could be from someone rude or making troubles out of nothing.

In this study, we use traditional machine learning models (i.e. Support Vector Machine) as baselines. We compare several deep learning models (i.e. Bidirectional Long Short Term Memory) with the baseline models to detect different types of online harassment such as threats, obscenity, insults, and identity-based hate. This textual online harassment scoring task is a multi-label multi-class classification problem and we build multiple One-vs-the-rest (OvR) classifiers for each type of online harassment classification. About machine learning experiment, we make use of a Wikipedia comments

corpus provided by the Conversation AI team[1], a research initiative founded by Jigsaw. Different textual representations (i.e. pre-trained word vectors GloVe) for the word embedding are used to extract textual features. Then we train our models on training set using cross-validation and test models on testing set. We use mean column-wise ROC AUC (Receiver Operating Characteristic Area Under Curve) as evaluation score metrics. Finally, we compare our proposed models with the baseline models to check whether our models will get better outcomes. About the econometric analysis, to quantify the economic impacts on the performance of crowdfunding projects, we examine two types of individual behavior using our proposed models: (1) textual online harassment of project visitors and (2) the attitudes of project creators towards textual online harassment. Our analysis is based on a large sample of 388,100 projects from 2014-04 to 2018-12, crawled monthly from Kickstarter, one of the most popular crowdfunding platforms.

This study contributes to the literature and casts lights on stakeholders of crowdfunding platforms in the following aspects: (1) As far as we know, this study is the first to propose a systematic framework for textual online harassment scoring from textual comments on crowdfunding platforms; (2) As far as we known, we are the first to examine to what extent the online harassment can account for project performance on crowdfunding platforms; (3) We assess the impact of the project creators' attitudes towards harassed comments on crowdfunding projects and the findings will provide important implications for reputation management of project and crowdfunding platform design.

Literature Review

Online harassment and incivility can take many specific forms, such as offensive speech, illegal harassment, social shaming, cyberbullying, and trolling (Lowry et al. 2016; Ransbotham et al. 2016). In this study, we put the emphasis on offensive speech (or hate speech), especially the toxic comment. Toxic comment is the textual form of offensive speech that are rude, disrespectful, hostile, aggressive or likely to make someone leave an online discussion, which differs from negative comments. Davidson et al. (2017) collect tweets containing hate speech keywords to build a crowd-sourced hate speech lexicon and trained a multi-class classifier to distinguish between different hate speech categories. And they found that racist and homophobic tweets are more likely to be classified as hate speech but that sexist tweets are generally classified as offensive. Gambäck and Sikdar (2017) designed a deep learning-based Twitter hate-speech text classification system assigning each tweet to one of four predefined categories: racism, sexism, both-hate-speech and non-hate-speech. Wulczyn et al. (2017) developed and illustrated a method combining crowdsourcing and machine learning methods to analyze personal attacks at scale and generated a corpus of over 100k high quality human-labeled comments from English Wikipedia, which is the basis of our training corpus.

Kickstarter is one of the most popular reward-based platforms, which utilizes the "all-or-nothing" crowdfunding mechanism. On Kickstarter, projects are divided into the following thirteen categories: Art, Comics, Dance, Design, Fashion, Film and Video, Food, Games, Music, Photography, Publishing, Technology, and Theater (Marom et al. 2016).

Textual Online Harassment Scoring: Toxic Comment Classification

Sometimes, distinguishing the frustrated grumbles from the malicious comments that are worthy of punishments is a difficult task. Using Wikipedia comments corpus with manual labels, we can make use of supervised learning techniques to deal with toxic comments. The textual online harassment score is between 0 and 1. The higher the textual online harassment score is, the higher tendency to online harassment the comment has. In our study, we focus on six types of textual online harassment : Toxic, Severe Toxic, Obscene, Threat, Insult and Identity Hate (Wulczyn et al. 2017). This textual online harassment scoring task can be regarded as a multi-label multi-class classification problem.

[1] Homepage: https://conversationai.github.io/

Baseline - Traditional Machine Learning Models

We use Multinomial Naive Bayes (MNB), Support Vector Machine (SVM), and Logistic Regression (LR) as the baseline models. Moreover, we follow the method proposed by Wang and Manning (2012) to develop Naïve Bayes versions of algorithms. Let $f^{(d)}$ be the frequency feature vector for document d with class label $y^{(d)} \in \{-1, +1\}$. This feature vector includes the frequency of each word w_i in document d. We define the frequency vector as $p = \alpha + \sum_{y^{(d)}=1} f^{(d)}$ and $q = \alpha + \sum_{y^{(d)}=-1} f^{(d)}$, where p and q are the frequency vector of two categories and α is the smoothing parameter. We take $\alpha = 1$ to for smoothing. Then, the log-count ratio is $r = log(\frac{p}{\|p\|_1} / \frac{q}{\|q\|_1})$. The feature vector $x^{(d)}$ is transformed to a new feature vector $\tilde{x}^{(d)} = x^{(d)} \circ r$. Finally, we can apply SVM and LR using the new feature vectors and get NB-SVM and NB-LR.

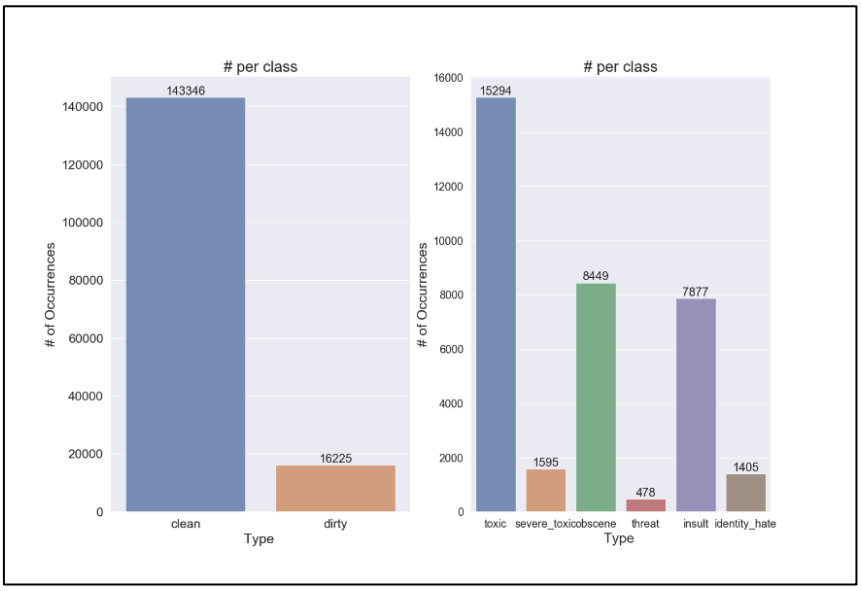

Figure 1. Histogram of clean and dirty comment

Deep Learning Models

Deep learning is a disruptive technique in many application scenarios and requires little feature engineering. In order to compare the deep learning models with traditional machine learning models, we make use of Long Short Term Memory (LSTM) model, Bidirectional LSTM (BiLSTM) and a hybrid of BiLSTM and convolutional neural network (CNN) in our study.

Long Short Term Memory: LSTM (Hochreiter and Schmidhuber 1997) is a novel recurrent neural network (RNN) and is capable of learning long-term dependencies. Comparing with traditional neural networks, RNN can persist information with loop and is suitable for sequence data, such as the textual comments in our study.

Bidirectional LSTM: Normal LSTM can only learn information from previous time steps rather than the future information. However, the future information is also important and can help us better understand the context. BiLSTM (Graves et al. 2013) contains two types of connecting sequences (forward and backward) and processes the data in both directions with two separate hidden layers.

BiLSTM-CNN: RNN allows embedding of information about sequences and previous words, while CNN can use this embedding to extract local features from them. A hybrid of Bidirectional LSTM and CNN (BiLSTM-CNN) is studied recently in some NLP tasks such as named entity recognition (Chiu and Nichols 2016). We add a CNN layer on the aforementioned BiLSTM architecture and check whether this hybrid model can improve the performance.

Machine Learning Experiment

Corpus

We will use a corpus containing a large number of Wikipedia comments labeled by human raters for toxic behaviour. Each comment was labeled by multiple annotators via *Crowdflower* who were asked whether the comment is toxic or healthy (Wulczyn et al. 2017). The training set containing 159,571 comment texts and six toxicity labels (*toxic*, *severe_toxic*, *obscene*, *threat*, *insult*, *identity_hate*) and the testing set containing 153,163 comment texts. True toxicity labels of the testing set are provided for the evaluation of models. In the training set, there are 143,346 comments that are not labeled as toxic comment ("clean" comments) and 16,225 comments that are labeled as toxic ("dirty" comments). Among the "dirty" comments, the number of each toxic label differs a lot (shown in Figure 1). There are 15,294 toxic, 8,449 obscene, 7,877 insult, 1,595 severe toxic, 1,405 identity hate and 478 threat comments.

Preprocessing

Text preprocessing is one of the most important parts of text classification, which can help us to feed more cleaned corpus to classifiers. We follow the standard text preprocessing process as the following description:

- We convert all the texts into lower case and remove some noisy and special characters such as linebreaks, usernames, IP address, URLs, email address, article IDs and so on.
- We tokenize each document d into words $[w_1, w_2, \ldots, w_n]$. In addition, among the tokenized words, we remove stop-words and English punctuations.
- We use *WordNet* from *NLTK* package for the Part-of-Speech (POS) tagging and word lemmatization.

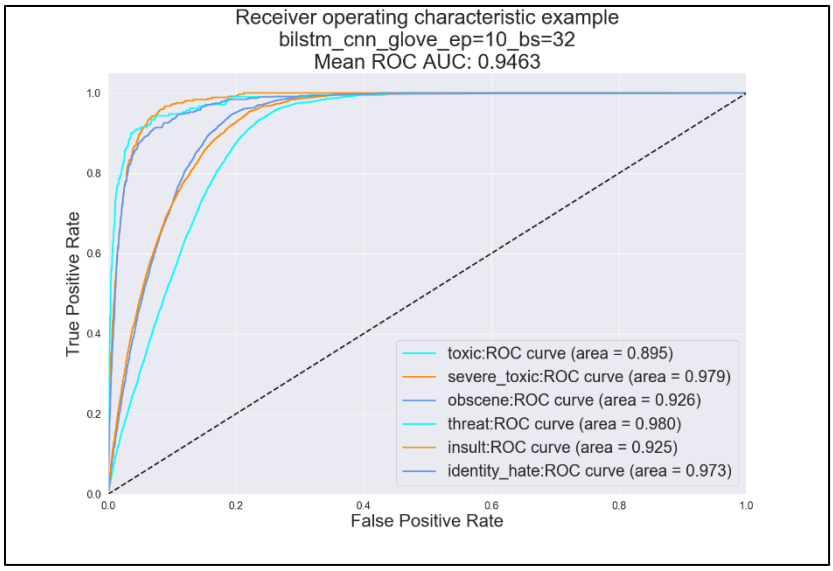

Figure 2. Plot of ROC curve. BiLSTM-CNN model with preprocessed corpus and pre-trained word vector GloVe. epoch = 10 and batch size = 32.

Textual Representation

Bag-of-words: We use "Bag-of-words" (BOW) representation in the traditional machine learning models. Documents are described by a vector of word occurrences, completely ignoring the relative positions of the words within the documents (Count vectorization). On the other hand, we can also re-weight the count features into token occurrence frequency using term frequency-inverse document

frequency (TF-IDF vectorization). Term frequency (TF) is the number of times a term occurs in a given document. And TF-IDF means term-frequency times inverse document-frequency: $tf - idf(t,d) = tf(t,d) \times idf(t)$. Here, $idf(t) = \frac{1+n_d}{1+df(d,t)} + 1$, where n_d is the total number of document, $df(d,t)$ is the number of documents containing term t.

Pre-trained Word Vectors: For deep learning models, text vectorization (also known as word embedding in deep learning) might be a little different and can be achieved via neural networks or matrix factorization. *word2vec* is one of the most popular ways and it transforms words into vectors, so that words with similar meaning end up laying close to each other. Similar to *word2vec*, *GloVe*[2] (Pennington et al. 2014), another well-recognized word representation method, is a count-based model. Both models are pre-trained on a large corpus to learn the co-occurrence information of words. As *GloVe* provides word vectors pre-trained on Wikipedia data, which is the same as our corpus, we utilize *GloVe* in our study.

Evaluation

We will use cross-validation for the hyperparameter tuning, which means that we will divide the training set into n folds, train algorithms on $n-1$ folds and test them on the left one fold for each iteration. For the evaluation of the models, we will use the mean column-wise ROC (Receiver Operating Characteristic) AUC (Area Under Curve) metrics (See Figure 2).

Sensitivity (True Positive Rate), Specificity (1-False Positive Rate), Precision and Recall are four related metrics to evaluate models. F1-score keep the balance between Precision and Recall and is given by the formula F1-score = 2 (Precision ×Recall)/(Precision + Recall). ROC curve illustrates the Sensitivity and (1-Specificity). The larger the ROC AUC, the better True Positives and True Negatives are distinguished. In a word, both F1 score and ROC AUC could be used to evaluate the performance of models. Given the class imbalance (shown in Figure 1), we prefer to use ROC AUC metric in our study. The mean column-wise ROC AUC score is the average of the individual AUCs of each predicted textual online harassment label.

Preliminary Results of Experiment

Baseline Models

In our study, MNB, SVM, LR, NB-SVM and NB-LR models are baseline models. For each baseline model, we make use of count vectorization scheme. First, we need to find the value of hyperparameter. Due to the time limit, we use a small sample data (20%) from the training/testing set to evaluate the hyperparameter. The preliminary results of baseline models can be found in Table 1.

Table 1. Preliminary Results - ROC AUC of different models

Model	Word Embedding	Best epoch & batch size)	Toxic	Serve Toxic	Obscene	Threat	Insult	Identity Hate	Mean
MNB	-	-	0.795	0.754	0.803	0.522	0.778	0.668	0.7199
SVM	-	-	0.776	0.674	0.776	0.772	0.738/	0.686	0.7368
LR	-	-	0.797	0.707	0.8	0.66	0.744	0.693	0.7334

[2] *GloVe* is an unsupervised learning algorithm for obtaining vector representations for words developed by Stanford NLP Group. Homepage: https://nlp.stanford.edu/projects/glove/

NB-SVM	-	-	0.777	0.694	0.767	0.785	0.721	0.685	0.7383
NB-LR	-	-	0.799	0.71	0.8	0.688	0.745	0.707	**0.7415**
LSTM	No GloVe	(10,32)	0.891	0.977	0.924	0.965	0.917	0.962	0.9393
LSTM	GloVe	(10,32)	0.892	0.978	0.924	0.979	0.920	0.968	0.9435
BiLSTM	No GloVe	(10,32)	0.891	0.978	0.927	0.940	0.917	0.930	0.9304
BiLSTM	GloVe	(10,64)	0.893	0.978	0.925	0.984	0.924	0.971	0.9458
BiLSTM-CNN	No GloVe	(10,32)	0.892	0.977	0.924	0.965	0.920	0.963	0.9400
BiLSTM-CNN	GloVe	(10,32)	0.895	0.979	0.926	0.980	0.925	0.973	**0.9463**

Deep Learning Models

We use *adam* optimizer (Kingma and Ba 2014) and binary cross-entropy function for the optimization. The word-embedding layer converts the word vectors into dense vectors of fixed size (50 in our study). We set the maximum volume of unique word features as 20,000 and maximum volume of highest weighted words in each comment as 100. The number of units in LSTM layer is 50. Due to the time limit, we can only tune the hyperparameters - batch size and epoch. We grid search the (epoch, batch size) pairs in (2,32), (2,64), (10,32) and (10,64). During training, we use 10-fold cross-validation. The preliminary results can be found in Table 1.

Summary

In Table 1, you can find that bidirectional LSTM outperforms baseline models and LSTM. Comparing with LSTM, BiLSTM can process the data in the forward and backward directions, which helps the model to better understand the context. Moreover, CNN can improve the performance of BiLSTM. As discussed above, CNN layer can help the model to extract local features, which can make the model outperform isolated BiLSTM model. In a word, the performance of deep learning models is much better than traditional machine learning models. The best baseline models is NB-LR model (ROC AUC: 0.7415), while the best deep learning model, BiLSTM-CNN with GloVe, can reach ROC AUC of 0.9360.

Considering the generality of Wikipedia comment corpus covering different topics and written in normative English, we can migrate the proposed methods to other corpora, such as crowdfunding platforms. Making use of the best deep learning model BiLSTM-CNN with GloVe, we can assess the online harassment propensity from English comments and use the extracted textual online harassment score for econometric analysis.

Econometric Methodology

Kickstarter Dataset

In this study, we use panel data from Kickstarter, which is crawled once a month from 2014-04 to 2018-12. In our sample, there are 388,100 projects from 22 countries covering 15 categories, among which 88.19% (342,275) are ended projects, which have reached their deadline for funding. Among the ended projects, 42.03% (143,870) are successfully funded. For the successfully funded projects, the total amount of project goals can reach 1.5 billion dollars and the total number of backers can reach 37.5 million. The average goal per project is 3,992.69 dollars and the average pledge per backer is 41.27 dollars. The average backers count per project is 96.75.

Econometric Model Specification

In this study, we investigate the impact of textual online harassment and the attitude of project creators towards the textual online harassment on the crowdfunding project performance. The dependent variable (response variable) is a binary variable illustrating the success or fail of projects. We can make use of survival analysis using a proportional hazards model to investigate the moderating effect of textual online harassment score and the creator's attitude towards textual online harassment (Kalbfleisch and Prentice 2011). As we are using panel data of Kickstarter, we can also make use of the quasi-experiment method, difference-in-difference (Donald and Lang 2007), to study the effects of factors we are interested in.

Following the crowdfunding literature (Etter et al. 2013; Mitra and Gilbert 2014; Mollick 2014), we adopt the following controls: (1) Project goal; (2) Project duration; (3) The number of pledge levels; (4) Minimum pledge amount; (5) Whether the project is featured in Kickstarter; (6) Whether the project has videos; (7) Whether the project has pictures; (8) Creator's Profile (Race, Gender, Country and so on); (9) Project Category; (10) The number of updates; (11) The number of comments; (12) The number of backers. Moreover, the econometric models that we are going to examine are described as follow:

Model 1: The independent variable is the textual online harassment scores for toxic comments.

Model 2: The independent variable is the creator's attitude towards textual online harassment including the number of toxic comments ignored, the number of toxic comment replied by the creator in a toxic way, the number of toxic comments replied by the creator in a polite way.

Discussion

In the machine learning experiment, parameter tuning is still not ideal due to the time limit. In the future, we can tune the hyperparameters of baseline models in a more fine-grained range and the hyperparameters of deep learning models other than epoch and batch size. Furthermore, we can add more layers for the deep learning models to check whether the performance can improve or not. Then, the F1 score metrics can be a supplementary evaluation for ROC AUC metric. We can also extract additional textual features from the corpus to help the classifier achieve better performance, such as the ratio of unique words in each comment. *word2vec* is predictive models, while GloVe is count-based models. And studies show that predictive models can achieve better performance than count-based models, such as Baroni et al. (2014). In the future, we hope we can improve our models via the aforementioned aspects.

As for econometric analysis, we are going to examine our proposed two models and check the impact of textual online harassment score and project creators' attitude towards textual online harassment. Finally, we can design experiments to compare the Kickstarter platform with one another well-known comparable crowdfunding platform such as Indiegogo.

Acknowledgements

We would like to acknowledge Conversation AI team from Jigsaw, Mr. SHEN Xinjia, and Prof. Antoni Chan from City University of Hong Kong here. This work is partially supported by research grants from Research Grants Council of Hong Kong (CityU 11504515; CityU 11508517), and strategic research grant (7004778) from City University of Hong Kong.

References

Akers, R. 2017. *Social Learning and Social Structure: A General Theory of Crime and Deviance*. Routledge.

Baroni, M., Dinu, G., and Kruszewski, G. 2014. "Don't Count, Predict! A Systematic Comparison of Context-Counting Vs. Context-Predicting Semantic Vectors," *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 238-247.

Chiu, J. P., and Nichols, E. 2016. "Named Entity Recognition with Bidirectional Lstm-Cnns," *Transactions of the Association for Computational Linguistics* (4), pp. 357-370.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. 2017. "Automated Hate Speech Detection and the Problem of Offensive Language," *Eleventh International AAAI Conference on Web and Social Media*.

Donald, S. G., and Lang, K. 2007. "Inference with Difference-in-Differences and Other Panel Data," *The review of Economics and Statistics* (89:2), pp. 221-233.

Etter, V., Grossglauser, M., and Thiran, P. 2013. "Launch Hard or Go Home! Predicting the Success of Kickstarter Campaigns," *Proceedings of the first ACM conference on Online Social Networks (COSN'13)*: ACM, pp. 177-182.

Gambäck, B., and Sikdar, U. K. 2017. "Using Convolutional Neural Networks to Classify Hate-Speech," *Proceedings of the first workshop on abusive language online*, pp. 85-90.

Graves, A., Jaitly, N., and Mohamed, A.-r. 2013. "Hybrid Speech Recognition with Deep Bidirectional Lstm," *2013 IEEE workshop on automatic speech recognition and understanding*: IEEE, pp. 273-278.

Hochreiter, S., and Schmidhuber, J. 1997. "Long Short-Term Memory," *Neural computation* (9:8), pp. 1735-1780.

Kalbfleisch, J. D., and Prentice, R. L. 2011. *The Statistical Analysis of Failure Time Data*. John Wiley & Sons.

Kingma, D. P., and Ba, J. 2014. "Adam: A Method for Stochastic Optimization," *arXiv preprint arXiv:1412.6980*).

Lowry, P. B., Zhang, J., Wang, C., and Siponen, M. 2016. "Why Do Adults Engage in Cyberbullying on Social Media? An Integration of Online Disinhibition and Deindividuation Effects with the Social Structure and Social Learning Model," *Information Systems Research* (27:4), pp. 962-986.

Maher, B. 2016. "Can a Video Game Company Tame Toxic Behaviour?," *Nature News* (531:7596), p. 568.

Marom, D., Robb, A., and Sade, O. 2016. "Gender Dynamics in Crowdfunding (Kickstarter): Evidence on Entrepreneurs, Investors, Deals and Taste-Based Discrimination," *Investors, Deals and Taste-Based Discrimination (February 23, 2016)*).

Mitra, T., and Gilbert, E. 2014. "The Language That Gets People to Give: Phrases That Predict Success on Kickstarter," *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*: ACM, pp. 49-61.

Mollick, E. 2014. "The Dynamics of Crowdfunding: An Exploratory Study," *Journal of business venturing* (29:1), pp. 1-16.

Pennington, J., Socher, R., and Manning, C. 2014. "Glove: Global Vectors for Word Representation," *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532-1543.

Ransbotham, S., Fichman, R. G., Gopal, R., and Gupta, A. 2016. "Special Section Introduction—Ubiquitous It and Digital Vulnerabilities," *Information Systems Research* (27:4), pp. 834-847.

Schwienbacher, A., and Larralde, B. 2010. "Crowdfunding of Small Entrepreneurial Ventures," *Handbook of entrepreneurial finance, Oxford University Press, Forthcoming*).

Wang, S., and Manning, C. D. 2012. "Baselines and Bigrams: Simple, Good Sentiment and Topic Classification," *Proceedings of the 50th annual meeting of the association for computational linguistics: Short papers-volume 2*: Association for Computational Linguistics, pp. 90-94.

Wulczyn, E., Thain, N., and Dixon, L. 2017. "Ex Machina: Personal Attacks Seen at Scale," *Proceedings of the 26th International Conference on World Wide Web*: International World Wide Web Conferences Steering Committee, pp. 1391-1399.