Automated Identification of Verbally Abusive Behaviors in Online Discussions

Srećko Joksimović
University of South Australia, Australia
srecko.joksimovic@unisa.edu.au

Ryan S. Baker
University of Pennsylvania, USA
rybaker@upenn.edu

Jaclyn Ocumpaugh
University of Pennsylvania, USA
jlocumpaugh@gmail.com

Juan Miguel L. Andres
University of Pennsylvania, USA
miglimjapandres@gmail.com

Ivan Tot
University of Defence in Belgrade, Serbia
Ivan.tot@va.mod.gov.rs

Elle Yuan Wang
Arizona State University, USA
elle.wang@asu.edu

Shane Dawson
University of South Australia, Australia
shane.dawson@unisa.edu.au

Abstract

Discussion forum participation represents a crucial support for learning and often the only way of supporting social interactions in online settings. However, learner behavior varies considerably in these forums, including positive behaviors such as sharing new ideas or asking thoughtful questions, but also verbally abusive behaviors, which could have disproportionate detrimental effects. To provide means for mitigating potential negative effects on course participation and learning, we developed an automated classifier for identifying communication that show linguistic patterns associated with hostility in online forums. In so doing, we employ several well-established automated text analysis tools and build on common practices for handling highly imbalanced datasets and reducing sensitivity to overfitting. Although still in its infancy, our approach shows promising results (AUC ROC=0.74) towards establishing a robust detector of abusive behaviors. We provide an overview of the classification (linguistic and contextual) features most indicative of online aggression.

1 Introduction

Massive Open Online Courses represent an important part of the educational landscape, offering access to learning at scale for both for-credit and life-long learners (Al-Imarah and Shields, 2019). While there is significant appeal and popularity in MOOC offerings, they bring numerous challenges for designing effective teaching and learning activities at scale (Kovanović et al., 2015). The unprecedented numbers of learners enrolled, and the diversity in learners' motivations and goals are but two factors that add a significant layer of complexity that is seldom experienced in more traditional modes of education (Carlos Alario-Hoyos et al., 2017). A product of the complexity of teaching at scale resides in the lack of student participation in discussion activity (Wise and Cui, 2018; Rosé and Ferschke, 2016). Despite social interactions between peers being a key factor in student learning (Poquet and Dawson, 2016; Joksimović et al., 2016), MOOC discussions often receive limited participation (Wise and Cui, 2018). Numerous studies have shown that participation in discussions is influenced by factors, such as feelings of confusion or isolation, diverse cultural and educational backgrounds, or the lack of ability to navigate when learning in a crowd (Baxter and Haycock, 2014; Poquet et al., 2018). Learners in MOOC settings require the rapid capacity to establish and sustain shared communication practices in order to join a new and often brief-lived online community (Rosé and Ferschke, 2016).

There is thus far relatively limited research on the pragmatics of academic discussions in MOOCs. In one line of work, surveys investigating why students stop posting in MOOC forums

Proceedings of the Third Workshop on Abusive Language Online, pages 36–45
Florence, Italy, August 1, 2019. ©2019 Association for Computational Linguistics

show that many quit because of comments deemed as politeness violations (Mak et al., 2010). Many of these postings involve relatively mild examples of abusive behaviors violations of pragmatic practices around niceness. More extreme violations of politeness conventions in MOOCs have also emerged in the literature, with Comer and her colleagues (Comer et al., 2015) reporting a number of verbally abusive behaviors on the part of students in MOOCs. While such behaviors are relatively infrequent, they can have disproportionate effects on those involved in the course (Mak et al., 2010; Comer et al., 2015).

In this work, we build on prior research on text classification and the analysis of learner generated discourse to build an automated classifier for detecting verbally abusive behaviors in online discussion forums. In so doing, we employ a wide variety of features that range from simple syntactic properties of text (such as unigrams, bigrams, or part-of-speech tags), to more complex linguistic analysis (e.g., text cohesion), in order to identify potentially relevant contextual features. We enhance these detectors through approaches designed to adjust for imbalance in data. The findings from this work bring new insights into the linguistic dimensions that could be indicative of online aggression that can help to mitigate the impacts of hostile and abusive behaviors on other learners.

2 Background Work

2.1 Roots of Negativity in MOOCs

Discourse around negativity in general, and MOOCs in particular, draws on the research on negative emotions in learning and use of abusive language in online learning communities (Comer et al., 2015). Experiencing anxiety, anger or frustration caused by learning activities that are being negatively valued or perceived as aversive, can lead to decreased engagement, motivation, and consequently failure to achieve specific learning outcomes (Pekrun et al., 2002; Rowe, 2017). On the other hand, with the emergence of social media and their use to support development of online learning communities, negativity and abusive online behaviors can potentially have much broader consequences (Salminen et al., 2018). Less extreme manifestations of abusive language in online learning communities could lead towards disengagement from the community (Mak et al., 2010). In more severe instances, negativity in online communities could lead to cyberbullying and online aggression in general (Holfeld and Grabe, 2012).

Designed to support interactions at scale and facilitated as a fully online learning experience, MOOCs pose multiple challenges to successful participation. For example, success in MOOCs is dependent on learners' motivation, achievement and social emotions, and self-regulatory learning skills (among other factors) (Mak et al., 2010). Therefore, as Rose and Ferschke (2016) posit, it is necessary to create "a supportive environment in which these learners can find community, support, dignity, and respect" (ibid., p664). In that sense, it seems reasonable to build on the approaches to mitigate abusive online behaviors commonly applied in online learning communities, then in more traditional educational settings.

To understand the nature of negativity in MOOCs, we draw on the work by Comer and her colleagues (2015) who discuss three types of negativity in MOOCs: negativity towards i) the course, ii) instructor, and iii) course platform. This multifaceted perspective demonstrates that the main sources of negativity are associated with pedagogy or course design decisions and cannot be easily addressed during course facilitation (Comer et al., 2015). Despite the relatively low proportion of abusive behaviors in MOOCs, Comer and colleagues illustrate the negative impacts they have on instructor presence and the broader levels of participation in discussion forums. Detecting when negativity occurs could provide the opportunity for a more automated or semi-automated approaches to reduce its impact, whether by blocking offensive content or deploying supportive strategies for the individuals impacted (Comer et al., 2015).

In this study we aim to automate the detection of negativity in MOOCs forums. An outcome of this work is to provide a process to enable more efficient responses to abusive online behaviors in MOOC discussion forums. In so doing, we treat negativity as a single construct, rather than differentiating negativity towards the course, platform, or instructor, due to the relative infrequency of negative behaviors. Although we concur that negativity in MOOCs can potentially have multiple facets, our goal in this study is to provide insight into factors that could indicate detrimental and abusive online behaviors in their broadest manifestation even negativity towards the course platform can be upsetting to others (Comer et al., 2015).

2.2 Automated Analysis of Abusive Language

Contemporary literature on affect in MOOC discourse primarily relies on content analysis methods (Joksimović et al., 2018b). To date, this has involved exploring affect and emotions to understand factors that predict persistence and success in MOOCs (Joksimović et al., 2018b). Tucker and colleagues (2014), for example, relied on a word-sentiment lexicon to extract sentiment polarity (i.e., positive, negative, or neutral) and strength (i.e., the magnitude of sentiment) from discussion forum messages. Tucker and colleagues found a strong negative association between the sentiment expressed in forums and average assignment grade. Adamopoluous (2013) opted for a more fine-grained analysis, exploring learners' sentiment towards course instructor, assignments, and course material, utilizing AlchemyAPI. Finally, Yang and colleagues (2015) relied on Linguistic Inquiry and Word Count (LIWC) features, and word categories that depict student affective processes, including positive and negative emotions. to detect confusion within student contributions to the discussion forum.

Although the existing MOOC research recognizes the importance of understanding learners' emotions expressed through interactions in online discussion forums, little has been done to detect negativity and abusive online behaviors. Relevant work exists, however, in efforts to understand online learning communities and social media interactions in general. Several approaches have been developed to detect dimensions of verbal aggression and abusive behavior in social media and online social platforms more broadly (Balci and Salah, 2015; Anzovino et al., 2018). For example, Abozinadah and Jones (2017) used Support Vector Machines (SVM) to detect abusive Twitter accounts. In another example, Anzovino and colleagues (2018), utilized a wide set of linguistic and bag-of-word features to explore the accuracy of various classifiers to identify misogynistic language on Twitter. The best classification accuracy was achieved using an SVM classifier based on unigrams, bigrams, and trigrams.

Additionally, a considerable body of research focuses on detecting verbal aggression in online social games, interactions with virtual partners, or the comments on popular news media (such as CNN.com or Yahoo! News) (Balci and Salah, 2015; Nobata et al., 2016). Relying on wide range of linguistic and contextual features (e.g., learner profile related information), Balci and Ali Salah (2015) used the Bayes Point Machine classification algorithm to identify online profiles that elicit abusive behaviors in social games. Nobata and colleagues (2016), on the other hand, explored the manifestation of abusive language in the comments posted on Yahoo! Finance and News articles. Nobata and colleagues (2016) developed a deep learning approach, utilizing n-grams, linguistic features (e.g., length of tokens, average length of word), syntactic features (e.g., par-of-speech tag of parent), and distributional semantics features.

Our work goes beyond existing approaches to understanding MOOC discourse, trying to detect abusive behaviors that could potentially have detrimental effects on teaching and learning. In so doing, we rely on features commonly identified as being predictive of learners' affective states and emotions in online learning settings. We also utilize algorithms and methods applied in general research on understanding verbal aggression in online learning communities in general.

3 Method

3.1 Data

The dataset for this study was obtained from the Big Data in Education MOOC, delivered from October to December 2013, by Columbia University, taught through the Coursera platform. This course iteration had a total of 45,256 enrolled learners during the course an additional 20,316 joined and accessed the course after its official end date. To successfully complete the course and receive a certificate, learners were required to earn an overall grade average of 70% or above. The overall grade was calculated by averaging the six highest grades extracted out of a total of eight assignments. All assignments were composed of multiple-choice questions and short numerical answers and as such, were available for automatic grading. Discussion participation was not graded. The majority of students only watched videos and did not participate in the assessment tasks. Some 1,380 students completed at least one assignment, while a total of 638 learners successfully completed the course.

Like vast majority of MOOC offerings, the discussion activity consists of a considerably small number of learners (Poquet and Dawson, 2016). For the MOOC under investigation, 747 unique users were engaged in discussion forum (*N*=747, including teaching staff). In total, the discussion forum contained 4,039 messages, written in English (*M*=5.41, *SD*=23.93). Two independent coders coded the dataset, labeling each message as being "negative", if at least one of the negativity types as defined by Comer and colleagues (2015) was found in a message, or "positive/neutral" otherwise. The process was performed through several phases. First 100 messages were analyzed together, to train the researchers and develop the coding scheme. After that, each of the coders independently labeled 200, 300, 400, and 500 messages, until a satisfactory percent agreement (%-agree = 96.6) was reached. The percent agreement was calculated at the end of each stage and all disagreements were discussed and resolved. The remaining messages (from 1,501 to 4,039) were split between the two coders.

Out of these 4,039 messages, 3,917 were positive/neutral, and 122 (3.02%) were coded as negative. From the total number of students who posted to discussion forum, 82 students posted at least one message coded as "negative" (*M*=1.49, *SD*=1.09). Nevertheless, only 9 students posted more than two messages coded as negative, showing repeated negativity towards the instructor, course platform, or course content.

3.2 Features

In order to develop a classification system for recognizing negativity in learners' posts in a discussion forum, we utilize several types of features. The extracted features build on those commonly used in the existing work on discourse analysis (Kovanović et al., 2014; Joksimović et al., 2014). Specifically, we rely on basic linguistic features (such as n-grams and part-of-speech tags), features extracted using tools for automated text analysis, and contextual features. The final feature set included 688 features.

3.2.1 Basic Linguistic Features

Our set includes some of the commonly used bag-of-words features, utilized in similar classification problems. Specifically, we extracted *n-gram features* (i.e., unigrams, bigrams, and trigrams), sequences of words that commonly appear together. Additionally, we extracted *part-of-speech tags* (e.g., noun, verb, adjective) and *syntactic dependency* (i.e., the relation between tokens) features. Although features like n-grams tend to inflate the feature space, these are often used as a baseline feature set, against which other features are compared to evaluate their contribution to the classification accuracy. Due to a limited training set size and unbalanced data, concerns about overfitting led us to use only the top most common 100 n-grams. All the basic features were extracted using Python programming language and the spaCy, open-source library for Natural Language Processing in Python.

3.2.2 Linguistic Facilities

In this study, we utilize three additional tools for advanced text analytics. Specifically, we use Linguistic Inquiry and Word Count (LIWC) to extract counts of different word categories, indicative of various psychological processes, such as social words, cognitive processes, or affect words (Tausczik and Pennebaker, 2010). Previous research demonstrates the potential of LIWC to capture different aspects of students' cognitive engagement during learning. For example, Kovanovic and colleagues (Kovanović et al., 2014), as well as Joksimovic and colleagues (Joksimović et al., 2014), showed that certain LIWC categories, such as the number of question marks or the number of first-person singular pronouns, are among the most important predictors of different phases of cognitive presence. Moreover, dimensions captured by LIWC (e.g., certainty, negations, or causal verbs), have been positively associated with (deactivating) negative emotions, such as boredom, anxiety, or frustration (D'Mello and Graesser, 2012).

We also utilize TAACO, a linguistic tool for automated analysis of text cohesion that provides more than 150 indicators of text coherence linguistic complexity, text readability, and lexical category use (Crossley et al., 2016). Dowell and colleagues (2015), and Joksimovic and colleagues (2018a), established the association between various metrics of text cohesion (e.g., referential or deep cohesion) and multiple social and academic learning outcomes. D'Mello and Graesser (2012), on the other hand, showed the association between cohesion-based metrics and student emotions (e.g.., boredom, engagement, confusion, or frustration) expressed during tutoring.

It seems also reasonable to expect that the negativity in discussion posts would be reflected through various emotional states. Therefore, we also used the IBM Watson Natural Language Understanding API to detect *anger*, *disgust*, *joy*, *fear*, and *sadness*, conveyed in discussion forum messages. Finally, given that research argues for the importance of considering sentiment expressed in discussion forums as being predictive of persistence in MOOCs, we extracted *sentiment polarity* and *sentiment subjectivity*, using TextBlob Python library for natural language processing tasks.

3.2.3 Contextual Features

Drawing on previous research by Kovanovic and colleagues (Kovanović et al., 2014), we further included contextual features into our feature space. As Comer and colleagues (2015) suggest, some of the learners posting negative messages in discussion forums tend to do so consistently. Therefore, for each post we observed whether the previous post by the same student was also negative. Moreover, it seems reasonable to expect that learners would build on the existing discourse, therefore we also observed whether there were negative messages in the same thread, prior to the observed post. Furthermore, we observed whether the posted message is a post or a comment, the start or the end of the thread, and number of votes the observed post received. Finally, for each of the posts we obtained an information whether the message contains positive and negative words, as well as the proportion of words that were positive and the proportion that were negative.

3.3 Model Implementation

We built our classifier using the Python scikit-learn implementation of Support Vector Machines (SVM), one of the most robust classifiers for text analysis (2014). In order to obtain optimal classification results, we performed hyperparameter optimization within the training set with parameters *C* (0.001, 0.01, 0.1, 1, 10) and *gamma* (0.001, 0.01, 0.1, 1), for each of the four kernels (i.e., “poly”, “rbf”, “linear”, “sigmoid”). We opted for the linear kernel, (*C*=0.001, *gamma*=0.001) as the settings with linear kernel yielded the best performance.

There are two challenges associated with the dataset that is inherent to the nature of the problem under study. Although the expression of negative or deactivating emotions is common within learning (Pekrun et al., 2002), verbally abusive behaviors are less common, although still detrimental (Mak et al., 2010; Comer et al., 2015). As indicated in our dataset, a small percentage of messages (3.02%) coded as “negative”, resulted in a highly imbalanced dataset, which could have negative effects on the classification results. In addition, participation in discussion forums, including the use of inappropriate or negative behaviors, varies by factors such as student demographics or motivation (Mak et al., 2010). Thus, the tendency to engage in inappropriate behaviors might (and does) vary from one learner to another. That is, only a small subset of students will express negativity in discussion forums.

To address the first problem of the highly imbalanced classes, we employed two strategies. First, the SVM classifier was configured to use balanced class weights. This configuration is used to adjust weights inversely proportional to class frequencies, defining higher weight for the “negative” class in our case. Second, we also implemented a False Positive Rate test into the classification pipeline. The False Positive Rate test controls for the total amount of false detections, which are common in imbalanced datasets with a rare category of interest, as in this study.

Cross-validation is typically used to control for overfitting. Desmarais and Baker (2012), highlight the importance of cross-validating at student level, to estimate goodness for new students rather than for new data from the same students. In our study, we rely on GroupKFold Python implementation of a K-fold iterator with non-overlapping groups (i.e., ensuring that each learner is only represented in a single fold).

4 Results

4.1 Model Training and Evaluation

Table 1 shows the results of our model selection and evaluation. To find the optimal model, we primarily rely on Area Under the Receiver Operating Characteristic Curve (ROC AUC) score, as Cohen’s statistics does not yield reliable estimates for highly imbalanced datasets, as it is the case in this study (Jeni et al., 2013). To obtain optimal results, we performed classification including various subsets of the original feature set (Table 1). The highest AUC ROC value with the complete feature set was 0.73 (SD=0.06). The clas-

Feature Set	Total Features	Class. Accuracy	F1 Score	ROC AUC
Baseline (Unigrams)	100	0.86	0.90	0.63
+ Ngrams	300	0.88	0.91	0.60
+ POS	397	0.88	0.91	0.60
+ TAACO	580	0.83	0.88	0.64
+ LIWC	673	0.84	0.89	0.65
+ Sentiment	680	0.84	0.89	0.65
+ Context	**688**	**0.86**	**0.90**	**0.73**
Unigrams + TAACO + LIWC + Sentiment + Context	303	**0.85**	**0.89**	**0.74**

Table 1: Classification results for different SVM configurations, varying the feature set used in predicting abusive language and p-value cutoff point at 0.05 for False Positive Rate test.

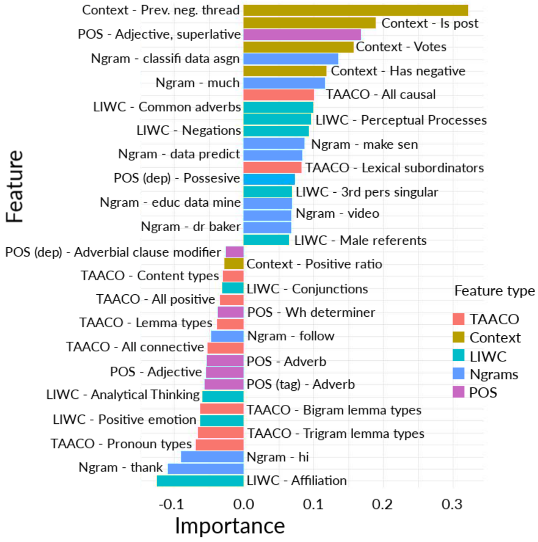

Figure 2: Top 40 features differentiating abusive language from overall positive/neutral language in discussion forum. It should be noted that values higher than 0 indicate features predictive of abusive language.

sification accuracy for the same set of parameters was .86 (SD=0.02), whereas the F1 score was .90 (SD=0.02).

Table 1 further shows that adding bigrams, trigrams and POS features (including tag and syntactic dependency) resulted in lower AUC ROC values, despite the slight increase in the classification accuracy. The ROC AUC score for the feature set that included Unigrams, TAACO, LIWC, Sentiment, and Contextual features was **0.74** (*SD*=0.06). The classification accuracy for the same set of parameters was .85 (*SD*=0.01), whereas the F1 score was .89 (*SD*=0.01).

4.2 Feature Importance Analysis

Given the size of the feature space (688 features), in the feature importance analysis we focus on the top 40 features used in the data separation task. That is, we observe the top 20 features most predictive of "negative" language and the top 20 features most predictive of "positive/neutral" language in the data set. Figure 2 shows that all groups of features (i.e., basic linguistic, features extracted using automated text analysis tools, and contextual features) are being identified within this subset of important features.

It is noteworthy that *contextual variables* yielded the highest predictive power for negativity (Figure 1). Specifically, `Previous negative thread` at least one of the previous messages in the thread was negative - has been identified as the most important variable in predicting detrimental behaviors. Moreover, whether a message is a post (i.e., reply to a thread) or a comment (i.e., reply to a post), as defined within the Coursera platform also revealed high predictive power. Finally, the total number of votes and whether message contained negative words were also found to be indicative of messages characteristic of negative behaviors towards the course content and design, course platform or course instructor.

Figure 1 further shows that part-of-speech tags representing `adjective in superlative` (e.g., "most", "worst"), were among the strongest predictors of negativity in online discussions. Other variables labeled as part of the part-of-speech dataset that were highly associated with negative messages are variables indicating the number of possession modifiers in a post (e.g., "... my experiences of the first hour in this class", "WASTE OF MY TIME"). On the other hand, variables indicative of positive/neutral messages were `adjectives`, `wh-determiners` (e.g., "what", "which"), and `adverbial clause modifiers` (e.g., "Confusion is good, just as long as it is addressed").

A considerable number of LIWC features were identified as being highly related to either negative or positive/neutral messages in MOOC discussions (Figure 1). Specifically, words associated with `common adverbs` (e.g., "write", "read", "hope"), `perceptual processes` (e.g., "watched", "said", "showed"), `negations` (e.g., "neither", "don't", "couldn't"), and function words that represent `3rd person singular form` (e.g., "him", "he's", "he"), were associated with messages indicative of abusive behaviors. On the other hand, words indicative of psychological processes representing core drives and needs (i.e., affiliation "welcome", "shared"), positive emotions (e.g., "helpful", "encourage", "honest"), analytical thinking, as well as function words (i.e., conjunctions "how", "then", "when"), were highly associated with positive/neutral behaviors (Figure 1).

Likewise, two variables extracted using TAACO linguistic facility were ranked among top 20 features predictive of "negative" messages. Specifically, count of causal connectives (e.g., "although", "because") and lexical subordinates (e.g., "unless", "whenever") were ranked as important variables in predicting abusive behavior. On the other hand, considerably more TAACO variables were identified as predictive of "positive/neutral" messages. Total number of content types, positive words, lemma types (including bigram and trigram lemmas), connectives, and pronoun types.

Several ngrams were also identified as important variables in differentiating abusive language from "positive/neutral" discourse. In the context of predicting "negative" messages, classify data assign, much, make sen, data predict, educ data mine, video, and dr baker emerged as the best predictors of abusive behaviors. Ngrams such as hi, thank, or follow, on the other hand, were associated with "positive/neutral" category of messages.

Observing variable importance with the smaller dataset (excluding part-of-speech, tag, and dependency variables) yielded rather similar results as the complete feature set (Figure 2). Contextual, LIWC, and ngrams (unigrams) still comprise a considerable part of the variables predictive of abusive behavior. Similarly, vide variety of TAACO variables was identified as indicative of "positive/neutral" messages.

5 Discussion and Conclusion

Identifying and mitigating abusive behaviors in the context of MOOCs is important for reducing the detrimental effects of negative language on peers and instructors. In this research, we manually coded all discussion forum messages written in English (N=4,039) from one MOOC, to build an automated classifier for identification of potentially harmful discussion messages. Our results show that primarily contextual, but also complex linguistic features, such as those extracted using LIWC and TAACO linguistic facilities represent important variables in predicting negativity in MOOCs. As such, our classifier outperforms, by a considerable margin, some of the recent work in identifying hate speech in online communities (Salminen et al., 2018).

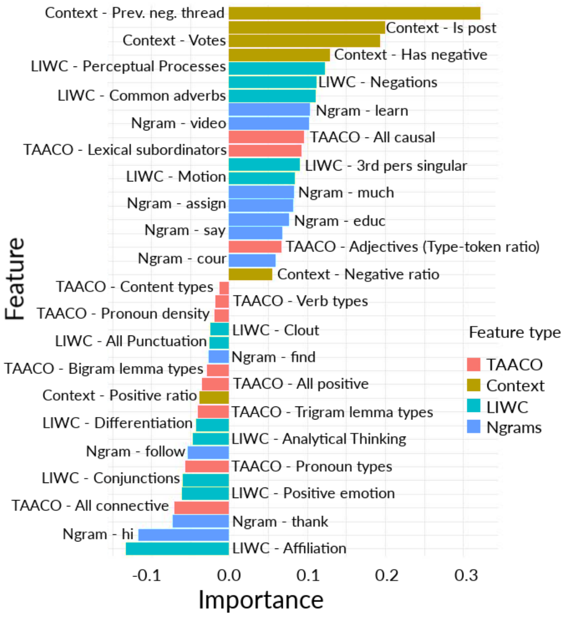

Figure 3: Features differentiating abusive language from overall positive/neutral language in discussion forum, for the model excluding bigram, trigram, and POS (including dependencies) features. It should be noted that values higher than 0 indicate features predictive of abusive language.

Kovanovic and colleagues (2014), argue for the importance of understanding the specific context in which certain messages in discussion forums have been posted. Our analysis on the complete and filtered feature set (without bigram, trigrams, and part-of-speech tag features) further support this finding. Moreover, the most important feature for predicting abusive language in MOOC discussions is a variable that flags whether the thread in which the current message has been posted already contains a "negative" message. This finding directly contributes to the claim made by Mak and colleagues (2010) or others, about the detrimental and likely disproportionate effect abusive language in MOOCs could have on the overall participation. The count of votes, as a contextual variable, also warrants further exploration. Complimenting others or content of others' messages represent one of the indicators identified within the social presence open communication category (Garrison and Akyol, 2013). However, one of the potential implications for future research could be exploration to what extent learners who express abusive behaviors in online communities tend to support each other. That is, to what extend acknowledgment and approval of negative behaviors implies negative connotation for the development of supportive learning environment and consequently learning success.

Our work also supports previous findings on understanding linguistic variables predictive of various dimensions of affect and emotions. For example, D'Mello and Graesser (2012) showed that the high ratio of causal words was positively associated with higher frustration. Whereas, negations were positively and significantly associated with boredom. Similar finding has been observed in our work where total count of all causal words was one of the main predictors of abusive language (Figure 2). Building further on Pekrun's (2002) control-value theory of achievement emotions, it seems that activities learners value negatively and perceive as not being controllable, potentially lead towards the abusive behaviors in online discussions.

It is also noteworthy that variables being identified as important predictors of "positive/neutral" messages, have been found to be associated with higher levels of cognitive engagement. For example, Joksimovic and colleagues [26] showed that the number of conjunctions (LIWC variable) or types of verbs (here captures using TAACO) were some of the variables positively and significantly associated with higher phases of cognitive inquiry, as defined by Garrison and colleagues [34]. This further supports the work by Rowe [13], among others, who showed that surface learners might be more likely to experience negative emotions, suggesting that "surface learners may react negatively to teaching methods which attempt to foster independent learning" (ibid., 299). Such a finding could have significant implications for future research and practice in mitigating abusive behaviors.

Although rather simple syntactic properties of text, such as ngram features, can easily inflate the feature space and result in overfitting, our results show that these variables should not be ignored. In the context of "negative" messages, it is indicative that unigrams, bigrams and trigram that emerged among the most important variables in predicting abusive behaviors, are related to specific aspects of the course (Figure 1 and 2). For example, ngrams such as "educ data mine", "video", "data predict", or "dr baker", indicate learners' focus on high level and general aspects of the course, rather than particular content related issues. On the other hand, among the most important variables in predicting positive/neutral messages, unigrams such as "hi" or "thank" emerged. Along with the LIWC variable "affiliation", these represent features indicative of higher levels of social presence [34]. Being recognized as important aspects of open and cohesive communication, as defined by Garrison and colleagues [34], these variables represent important indicators of tendency to establish collaborative and engaging community of learners.

5.1 Limitations

Although the dataset is reasonably large among text classification problems, high data imbalance represents one of the main challenges to this study. Moreover, in this preliminary analysis, we rely on the dataset from a single, technical MOOC (i.e., focused on the topics of big data and statistics). Future work should account for different subject domains and different educational settings (e.g., more formal traditional online courses).

References

Ehab A. Abozinadah and James H. Jones, Jr. 2017. A Statistical Learning Approach to Detect Abusive Twitter Accounts. In *Proceedings of the International Conference on Compute and Data Analysis*, ICCDA '17, pages 6–13. ACM.

Panagiotis Adamopoulos. 2013. What Makes a Great MOOC? An Interdisciplinary Analysis of Student Retention in Online Courses. In *34th International Conference on Information Systems*, United States. Association for Information Systems. The most heavily-cited paper from the ICIS 2013 proceedings (as of August 15th, 2016).

Ahmed A. Al-Imarah and Robin Shields. 2019. MOOCs, Disruptive Innovation and the Future of Gigher Education: A Conceptual analysis. *Innovations in Education and Teaching International*, 56(3):258–269.

M. Anzovino, E. Fersini, and P. Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10859 LNCS:57–64.

Koray Balci and Albert Ali Salah. 2015. Automatic Analysis and Identification of Verbal Aggression and Abusive Behaviors for Online Social Games. *Computers in Human Behavior*, 53:517 – 526.

Jacqueline Baxter and Jo Haycock. 2014. Roles and Student Identities in Online Large Course Forums: Implications for Practice. *The International Review of Research in Open and Distributed Learning*, 15(1).

Carlos Alario-Hoyos, Iria Estévez-Ayres, Mar Pérez-Sanagustín, Carlos Delgado Kloos, and Carmen Fernández-Panadero. 2017. Understanding Learners Motivation and Learning Strategies in MOOCs. *The International Review of Research in Open and Distributed Learning*, 18(3).

Denise Comer, Ryan Baker, and Wang Yuan. 2015. Negativity in Massive Online Open Courses: Impacts on Learning and Teaching and How Instructional Teams May Be Able to Address It. *InSight: A Journal of Scholarly Teaching*, 10:92 – 113.

Scott A Crossley, Kristopher Kyle, and Danielle S McNamara. 2016. The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior research methods*, 48(4):1227–1237.

Michel C. Desmarais and Ryan S. J. d. Baker. 2012. A Review of Recent Advances in Learner and Skill Modeling in Intelligent Learning Environments. *User Modeling and User-Adapted Interaction*, 22(1):9–38.

S. K. D'Mello and A. Graesser. 2012. Language and discourse are powerful signals of student emotions during tutoring. *IEEE Transactions on Learning Technologies*, 5(4):304–317.

Nia M Dowell, Oleksandra Skrypnyk, Srećko Joksimović, Arthur C Graesser, Shane Dawson, Dragan Gašević, Thieme A Hennis, Pieter de Vries, and Vitomir Kovanović. 2015. Modeling Learners' Social Centrality and Performance through Language and Discourse. *International Educational Data Mining Society, Paper presented at the 8th International Conference on Educational Data Mining (EDM) (8th, Madrid, Spain, Jun 26-29, 2015)*.

Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. 2014. Do We Need Hundreds of Classifiers to Solve Real World Classification Problems? *The Journal of Machine Learning Research*, 15(1):3133–3181.

D Randy Garrison and Zehra Akyol. 2013. The Community of Inquiry Theoretical Framework. In *Handbook of distance education*, pages 122–138. Routledge.

Brett Holfeld and Mark Grabe. 2012. Middle School Students' Perceptions of and Responses to Cyber Bullying. *Journal of Educational Computing Research*, 46(4):395–413.

László A Jeni, Jeffrey F Cohn, and Fernando De La Torre. 2013. Facing Imbalanced Data Recommendations for the Use of Performance Metrics. *International Conference on Affective Computing and Intelligent Interaction and workshops : [proceedings]. ACII (Conference)*, 2013:245–251.

Srećko Joksimović, Nia Dowell, Oleksandra Poquet, Vitomir Kovanović, Dragan Gašević, Shane Dawson, and Arthur C. Graesser. 2018a. Exploring Development of Social Capital in a cMOOC Through Language and Discourse. *The Internet and Higher Education*, 36:54 – 64.

Srećko Joksimović, Dragan Gašević, Vitomir Kovanović, Olusola Adesope, and Marek Hatala. 2014. Psychological Characteristics in Cognitive Presence of Communities of Inquiry: A Linguistic Analysis of Online Discussions. *The Internet and Higher Education*, 22:1 – 10.

Srećko Joksimović, Areti Manataki, Dragan Gašević, Shane Dawson, Vitomir Kovanović, and Inés Friss de Kereki. 2016. Translating network position into performance: Importance of centrality in different network configurations. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge (LAK'16)*, LAK '16, pages 314–323, New York, NY, USA. ACM.

Srećko Joksimović, Oleksandra Poquet, Vitomir Kovanović, Nia Dowell, Caitlin Mills, Dragan Gašević, Shane Dawson, Arthur C. Graesser, and Christopher Brooks. 2018b. How do we model learning at scale? a systematic review of research on moocs. *Review of Educational Research*, 88(1):43–86.

Vitomir Kovanović, Srećko Joksimović, Dragan Gašević, and Marek Hatala. 2014. Automated cognitive presence detection in online discussion transcripts. In *Proceedings of the Workshops at the LAK 2014 Conference co-located with 4th International Conference on Learning Analytics and Knowledge (LAK'14)*, Indianapolis, IN.

Vitomir Kovanović, Srećko Joksimović, Dragan Gašević, George Siemens, and Marek Hatala. 2015. What public media reveals about MOOCs: A systematic analysis of news reports. *British Journal of Educational Technology*, 46(3):510–527.

Sui Mak, Roy Williams, and Jenny Mackness. 2010. Blogs and forums as communication and learning tools in a MOOC. In *Proceedings of the 7th International Conference on Networked Learning 2010*, pages 275–285. University of Lancaster.

C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang. 2016. Abusive language detection in online user content. In *25th International World Wide Web Conference, WWW 2016*, pages 145–153.

Reinhard Pekrun, Thomas Goetz, Wolfram Titz, and Raymond P. Perry. 2002. Academic emotions in students' self-regulated learning and achievement:

A program of qualitative and quantitative research. *Educational Psychologist*, 37(2):91–105.

Oleksandra Poquet and Shane Dawson. 2016. Untangling MOOC learner networks. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, LAK '16, pages 208–212. ACM.

Oleksandra Poquet, Nia Dowell, Christopher Brooks, and Shane Dawson. 2018. Are MOOC forums changing? In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, LAK '18, pages 340–349. ACM. Event-place: Sydney, New South Wales, Australia.

Carolyn Penstein Rosé and Oliver Ferschke. 2016. Technology Support for Discussion Based Learning: From Computer Supported Collaborative Learning to the Future of Massive Open Online Courses. *International Journal of Artificial Intelligence in Education*, 26(2):660–678.

Anna D. Rowe. 2017. *Feelings about feedback: the role of emotions in assessment for learning*, The Enabling power of assessment, pages 159–172. Springer, Springer Nature, United States.

Joni Salminen, Hind Almerekhi, Milica Milenkovi, Soon gyo Jung, Jisun An, Haewoon Kwak, and Bernard Jansen. 2018. Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media. In *International AAAI Conference on Web and Social Media*, pages 330–339.

Yla R. Tausczik and James W. Pennebaker. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Conrad Tucker, Barton K. Pursel, and Anna Divinsky. 2014. Mining Student-Generated Textual Data In MOOCS and Quantifying Their Effects on Student Performance and Learning Outcomes. In *2014 ASEE Annual Conference & Exposition*, Indianapolis, Indiana. ASEE Conferences. Https://peer.asee.org/22840.

Alyssa Friend Wise and Yi Cui. 2018. Unpacking the relationship between discussion forum participation and learning in MOOCs: Content is key. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, LAK '18, pages 330–339. ACM. Event-place: Sydney, New South Wales, Australia.

Diyi Yang, Miaomiao Wen, Iris Howley, Robert Kraut, and Carolyn Rose. 2015. Exploring the effect of confusion in discussion forums of massive open online courses. In *Proceedings of the Second (2015) ACM Conference on Learning @ Scale*, L@S '15, pages 121–130, New York, NY, USA. ACM.