Stacked Generalization Learning to Analyze Teenage Distress

Karthik Dinakar[†] **Emily Weinstein**[*] **Henry Lieberman**[†] **Robert Selman**[*]

[†]Massachusetts Institute of Technology, [*]Harvard University
{karthik,lieber}@media.mit.edu {emily weinstein@mail, selmanro@gse}@harvard.edu

Abstract

The internet has become a resource for adolescents who are distressed by social and emotional problems. Social network analysis can provide new opportunities for helping people seeking support online, but only if we understand the salient issues that are highly relevant to participants personal circumstances. In this paper, we present a stacked generalization modeling approach to analyze an online community supporting adolescents under duress. While traditional predictive supervised methods rely on robust hand-crafted feature space engineering, mixed initiative semi-supervised topic models are often better at extracting high-level themes that go beyond such feature spaces. We present a strategy that combines the strengths of both these types of models inspired by Prevention Science approaches which deals with the identification and amelioration of risk factors that predict to psychological, psychosocial, and psychiatric disorders within and across populations (in our case teenagers) rather than treat them post-facto. In this study, prevention scientists used a social science thematic analytic approach to code stories according to a fine-grained analysis of salient social, developmental or psychological themes they deemed relevant, and these are then analyzed by a society of models. We show that a stacked generalization of such an ensemble fares better than individual binary predictive models.

Introduction

That there is a substantial amount of research documenting the negative effects of adolescent internet use such as cyber-bullying and sexual predators is well known (Gross and Acquisti 2005; Bogdanova, Rosso, and Solorio 2012; Guan and Huck 2012). Yet there is also research that has shown that distressed teenagers seek help and advice anonymously on the internet for a plethora of issues ranging from social and romantic relationships to self-injurious behavior to sexuality (Suzuki and Calzo 2004). This presents new opportunities for collaboration between the fields of prevention science and machine learning to a) understand the salient issues behind adolescent distress on a large scale and b) develop reflective user interfaces that provide help to participants by paying attention to their individual circumstances.

In this paper, we analyze posts from a popular teen support network through the lens of prevention science and discuss how this analysis could be practically deployed on the network to help its participants. We use findings from this exercise to present a computational framework consisting of a society of supervised and unsupervised models to extract the most dominant themes behind teenagers' stories. The contributions of this paper are threefold. First, we show that a stacked generalization of an ensemble of models for topic prediction works better than the individual models. Second, our approach highlights the importance of fine-grain analysis of the data by prevention science experts and a contextual inquiry behind using this analysis to power a reflective user interface prior to modeling. Third, we use our computational framework to produce a landscape of themes impacting distressed adolescents on this network.

Background and related work

The social network *A Thin Line* launched by MTV (Music Television) in 2010 is a website designed to help distressed teenagers with issues ranging from digital abuse to bullying and sexting. The website offers information and advice on how a teen might cope with such issues and encourages them to share their stories publicly in the hope that crowd-sourced responses and feedback to these stories by visitors to the website might be of some help to the distressed teenagers who share them. Similar efforts such the popular advice forum *TeenHelp.Org*, *Nerve Dating Confession* etc., have sprung in recent years to support distressed teenagers, providing an unvarnished glimpse into the world of adolescent issues.

Related work in this area can be examined from two broad perspectives, namely prevention science and the computational linguistics and human-computer interaction communities in computer science. In prevention science, recent work has focused on how anonymity and support groups help distressed teenagers (Webb, Burns, and Collin 2008) and on ethnographic studies examining teenage drama (Boyd and Marwick 2011). Studies have also focused on the growth of interpersonal understanding (Selman 1980), friendship in youth and pair-therapy (Selman and Schultz

Age	Poster is female	Poster is male	Gender n/a	Over the line	On the line	Under the line
Less than 10	4	0	0	1	0	0
10 to 15	2059	222	44	1611	2	133
16 to 20	1474	276	28	1283	9	79
21 to 25	135	61	2	138	0	8
Above 25	62	47	4	80	0	12
n/a	105	7	2617	1840	8	154

Figure 1: **The ATL dataset**: A summary of 7147 stories in *A Thin Line* dataset. A majority of the stories are in the age-group of 12 to 17 for both genders. The distribution of ratings is dominated by the rating 'over the line', indicating the severity of the personal accounts shared on the website.

1998), the role of isolation in adolescent depression and the importance of self-compassion and social support for well-being. (Neff and McGehee 2010). In computer science, kernel-based supervised learning methods and topic modeling have been used to model textual cyber-bullying (Dinakar, Reichart, and Lieberman 2011; Dinakar et al. 2012b; 2012a) in the machine learning community and the study of support groups and networks(Doherty, Coyle, and Sharry 2012; Farzan et al. 2012) in the CHI community, as well as predicting depression on social media (Choudhury et al. 2013).

Dataset

On the *A Thin Line* (ATL) platform, teenagers are encouraged or warranted to share a personal story in 250 characters or less. Once a story is posted on the site, visitors can rate a story under three categories: an *over the line* rating for stories that are deemed serious, *under the line* for stories that are deemed harmless, and an *on the line* for stories about which commenters are uncertain or ambivalent as to its seriousness. We obtained a set of 7,144 stories (along with their ratings, comments, the age and gender of posters) posted on this site over a period of three years from 2010 to 2013. The age and gender fields are self-reported and were not present for every poster.The dataset, which contains no personally identifiable information of its participants, was obtained through a licensing agreement with Viacom (MTV's parent company). Of the 7,144 personal accounts posted to the site, 4,415 (61.8%) included age (mean age= 16.3 years; sd = 5.21 modal user, according to self-reports of age and gender, is 15-years-old and female) and 4,466 (62.5%) included gender (86.3% reported they are female). The modal user, according to self-reports of age and gender, is 15-years-old and female.

Understanding the nuances in the dataset:a prevention science approach

Prevention Science as used within the disciplines of psychology and psychiatry investigates the causes and correlates of psychological dysfunctions with focus on preventing them. It identifies a set of 'risk' factors that lead to such a dysfunction. Prevention research is the systematic study of precursors or *"risk factors"* that lead to, contribute to, or predict to a psychosocial problem as well the search for protective factors or attributes that minimize the onset of a problem in the presence of risk factors. (Mrazek and Haggerty 1994; Selman and Dray 2006) Findings from prevention science inform policy and psychological methods to boost such *protective factors* and counter risk and investigate ways of preventing such problems from arising (Coie et al. 1993) and minimizing the damage when they do.

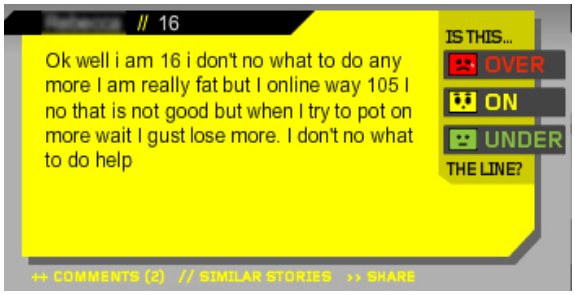
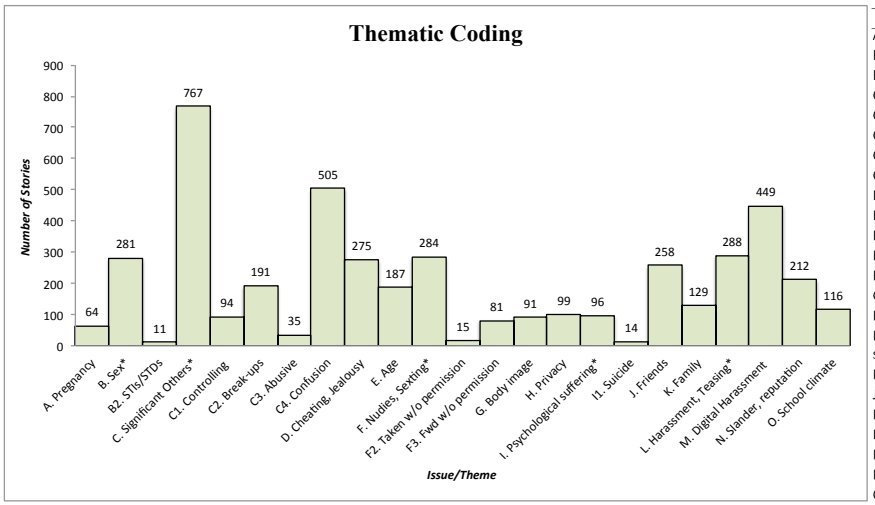


Figure 2: **A sample story from a participant on ATL**: each story in the dataset is restricted to not more than 250 characters. The figure also shows affordances for rating this story according to its severity - an *over the line* rating for stories that are deemed serious, *under the line* for stories that are deemed harmless, and an *on the line* for stories about which commenters are uncertain or ambivalent as to its seriousness.

Given that adolescents today spend an ever increasing proportion of their time using digital technologies (Madden et al. 2013), there is an unprecedented opportunity to understand the plethora of challenges they face on a big scale. Adopting a prevention science approach to support teenagers in the face of such challenges is conditioned on understanding the issues or stressors they face. With this



	total # stories
A. Pregnancy	64
B. Sex*	281
B2. STIs/STDs	11
C. Significant Others*	767
C1. Controlling	94
C2. Break-ups	191
C3. Abusive	35
C4. Confusion	505
D. Cheating, Jealousy	275
E. Age	187
F. Nudies, Sexting*	284
F2. Taken w/o permission	15
F3. Fwd w/o permission	81
G. Body image	91
H. Privacy	99
I. Psychological suffering*	96
I1. Suicide	14
J. Friends	258
K. Family	129
L. Harassment, Teasing*	288
M. Digital Harassment	449
N. Slander, reputation	212
O. School climate	116

Figure 3: **Emic coding**: Results of the emic thematic coding exercise over a sample of 2000/7147 stories. The most frequent theme in the above distribution is one that involves significant others. Romantic relationships {**C, C1, C2, C3, C4, D, B, F, F2, F3, A**} was the single biggest theme contributing to over 70% of the dataset. Each story was allowed to have more than one code. (*Includes count of sub codes). Three key aspects are worth noting in this coding scheme: **(1)** some of the codes represent distressing events, such as **C3** (abuse) and **L** (harassment and teasing), while some of them represent emotional reactions, such as **D** (jealousy) and **I** (psychological suffering), **(2)** some of the codes have relatively clearer and smaller repertoire of verbiage (age, sex etc) whereas abstract notions such as jealousy and controlling relationships are harder to map to specific verbiage, and **(3)** the codes and sub codes were developed independent of the notion that they would be consumed computationally.

context, we analyze the ATL dataset with two questions - a) What is the distribution of themes or issues raised by the personal stories in the dataset and b) What is the best way of translating the output of this analysis via computational models on a large scale? We adopt an interdisciplinary mixed-initiative approach towards a systematic identification of salient themes in the stories using a qualitative (emic) technique known as **thematic coding** as well as the iterative translation of these themes into a computational predictive framework.

Thematic emic coding

The emic approach (Boyatzis 1998) to analyzing a dataset is an *inductive* and *bottom-up* technique that in this case sets as its starting point the perspectives expressed by the participants in the dataset. In adopting an emic approach, a researcher sets aside preconceived assumptions and biases as best as possible to let the data speak for itself as it is analyzed. The themes, patterns and concepts derived from the dataset are generated through a systematic, fine-grain analysis. A related, but opposite approach in psychology is an *etic* approach which is more *deductive* and *top-down*, that has as its starting point theories and concepts beyond the setting from which the dataset was generated. We adopt an emic approach to study the ATL dataset to produce a *codebook* (Weinstein and Selman under review) in three analytical stages as follows:

Step 1: All stories were binned into one of four buckets based on the severity ratings assigned by the visitors to the website, where the majority threshold crossed 50% for a given story (i.e., a majority rated the story as over the line, a majority rated the story as under the line, a majority rated the story as on the line); stories with no majority rating were assigned to a fourth bucket.

Step 2: 200 stories were randomly selected from each of the buckets and the emic thematic coding process began to unearth the breadth and prevalence of the issues and codes assigned to topical issues (in the machine language sense) or themes (in the emic qualitative methods sense) in the stories by two prevention science experts. 23 thematic codes were ultimately generated (for 23 separate topics), with patterns (inclusions) and anti-patterns (exclusions) for each thematic code. A random sample of 100 stories selected from each bucket was blind coded for a multi-step, inter-rater reliability process, to clarify and sharpen the definition of each thematic code. After two rounds of thematic coding, a respectable kappa statistic (Viera, Garrett, and others 2005) (ranges 0.74 to 1) was achieved for each each code.

Step 3: The codebook derived from the previous step was applied to a total of 2000 out of the original 7147 stories. Each story could have more than one thematic code assigned to it.

For example, consider the following story (personal account):

"So my ex-boyfriend got me pregnant and cheated on me with my BFF now she's pregnant our due dates are one month apart and now he's with our other friend. He isn't claiming either of our babies and says he was never with us

& she's denying knowing us."

The thematic codes assigned to this one story were {**A, D, J**} for **pregnancy, cheating and involving friends** respectively. Our next step was to transition to machine learning [etc], using the thematic codes surfaced through the social science coding and the 2,000 expert-coded stories to aid in the development of reliable computational models that can then be applied to the remaining 5147 stories so as to plot the distribution of these codes across the entire ATL dataset.

Predictive models

We begin the process of building models to predict the codes for a given story with supervised learning methods widely used for text classification. This is an instance of multi-label classification, where each story could have more than one class label (or code) present in it. Given the relatively small size of the coded dataset for certain codes and the multi-label problem formulation, a popular line of research in statistical NLP has shown that combining individual binary classifiers for each label fares far better than a single multi-label predictive model (Viera, Garrett, and others 2005; Dinakar et al. 2012b).

Base Binary Models

We begin by training an ensemble of base models for predicting individual labels, namely a support-vector machine with a linear kernel (SVM-L), a radial basis function kernel (SVM-R) and a stochastic gradient boosted decision trees (GBDT) model (Guyon, Boser, and Vapnik 1993; Friedman 2001). We use the package scikit-learn(Pedregosa et al. 2011) for the gradient boosted decision trees.The coded dataset was split into training (70%), validation(15%) and test(15%) sets maintaining the proportion of labels in each the same as the original dataset. We adopted the following process to train the ensemble of base classifiers for each label as follows.

Data preprocessing Each story was subjected to preprocessing involving the removal of the mean value of a feature and dividing non-frequent features by their standard deviation after removal of stop words.

Feature Engineering We used a 10-fold cross validation for feature space engineering. Features for each label consisted of (1) unigrams, (2) bigrams and (3) bigram part-of-speech tags based on the Stanford POS tagger (Toutanova et al. 2003) and *tf-idf* features. An exhaustive list of these features was filtered using a chi-squared test and was further modified by adding or removing features to boost accuracy or to avoid over fitting iteratively. Each iteration of feature engineering was validated against the validation set. The feature set for each label was validated for each of the three base classifiers, SVM-L, SVM-R and GBDT.

Parameter sweeps & model selection The finalized set of features were then subjected to an exhaustive parameter sweep to estimate the optimal hyper-parameters for all the three types of classifiers, for each label. Given 23 codes, there were 69 parameter sweeps in total, one for each label under each category of classifiers. Figure 4 shows a summary of the parameter sweep for all three classifiers against the test set. The best model with respect to F1 score was selected for each code. Despite hyper-parameter fine-tuning and an iterative refinement of the feature space, it can be clearly seen that the performance of the base models for certain codes (example **F3**, **C3**, **C**) still remained low.

Error Analysis An error analysis for each base category revealed two issues: a) classifiers for codes with low F1-score (F-measure) could have done better with features indicating the presence of related codes in the story and b) there were patterns of code dominance, where the dominance of a specific code indicated the co-presence of another. This was especially true for codes that had low F1-score (F-measure). For example, consider the following two stories in the test set:

T1:*"the guy i lost my virginity to secretly taped us on his webcam and i wasn't told until a couple months later when all of his friends asked me how i felt to be the mini pornstar of our small town school. i dont know what to do. he doesnt know i know."*

T2:*"Me and my bf have been on and off for almost 3 years now. I messed up badly and cheated on him. and now our relationship has turend emotionally abusive. He calls me inappropriate names like a **** and a *** and tell me to go *** a ***. whatdoido?*

In personal account**T1**, the actual labels assigned were {**F3,F2,N**}, whereas the base model framework missed **F2**. In personal account **T2**, the actual labels assigned were {**D, C3, L, C4**} but the model missed **C3**. A re-examination of the training set revealed that 32% of stories with multi-codes with N as one of them also had **F2** as a co-occurring code; stories with **D** as the most dominant theme (as interpreted by the prevention science experts) also had **C3** as a co-occurring code, thereby necessitating the capturing of co-occurring themes as well as code dominance.

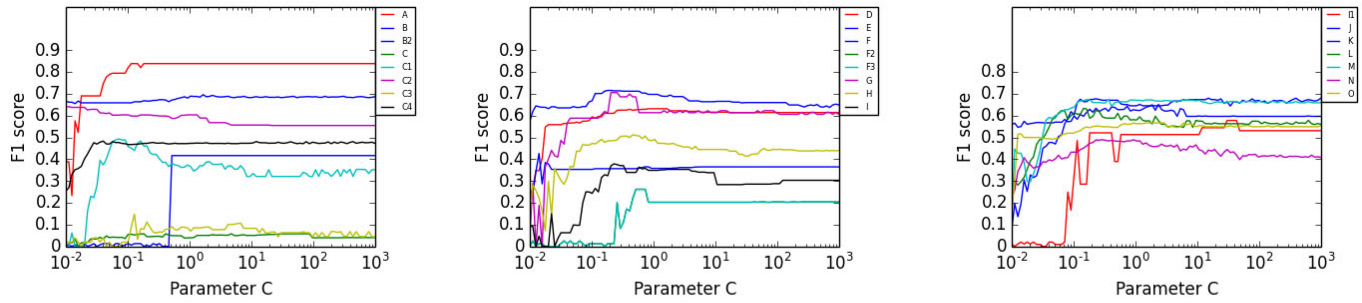
Capturing code proportion & co-occurrence

One way of capturing code co-occurrence and proportion is through a topic model. Given a multi-labeled dataset, a natural choice to model this is labeled latent Dirichlet Allocation (L-LDA), which has shown to perform well in such a scenario (Ramage et al. 2009) and on microblogs with documents of a short length (Ramage, Dumais, and Liebling 2010).

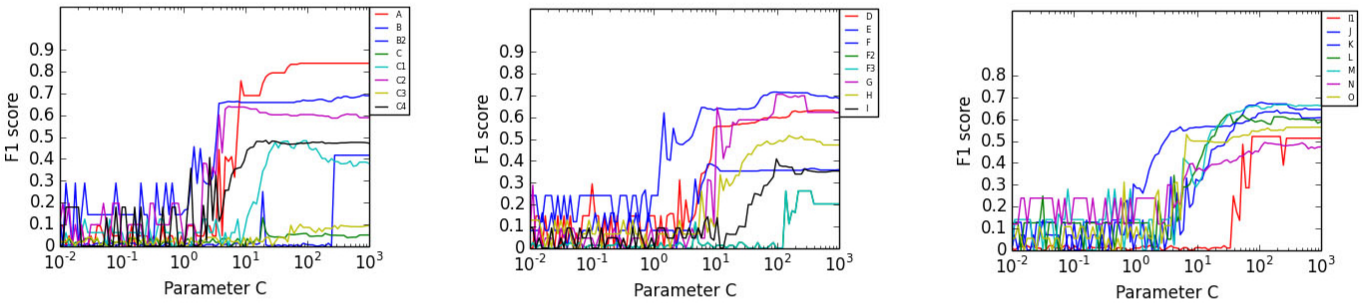
We trained a labeled LDA model based on the paper by Ramage (Ramage et al. 2009) using the training and validation set splits generated for the base classifiers. Model selection involved varying the hyper-parameters α and β measured against the validation set. Estimates for the hyper parameters after parameter sweep were $\alpha = 0.03$ and $\beta = 0.01$.

Error Analysis An error analysis against the test set underlines the limitations of the discriminatory power of semi-supervised topic models. New stories that have in them new phrases and verbiage not present in the vocabulary set of the training corpus lead to misclassification. While a discriminatory model like a support-vector machine's feature spaces

(a) SVM-L: model selection of vector machines with linear kernel classifiers, parameter sweeps of soft-margin C against F1 score



(b) SVM-R: model selection of vector machines with radial basis kernel classifiers, parameter sweeps of soft-margin C against F1 score



(c) GBDT: model selection of gradient boosted decision tree classifiers for each code, parameter sweeps of # of estimators against F1 score

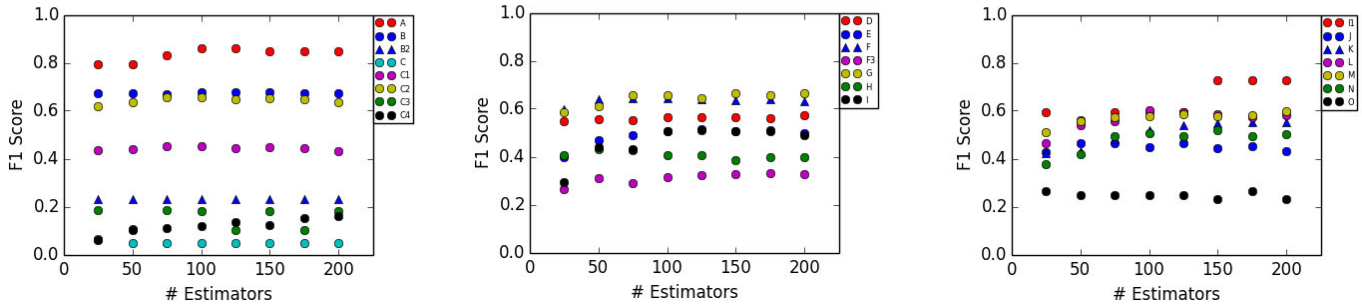


Figure 4: **Need for a meta-learner:** The above figures shows results from the parameter sweeps. Each row shows a parameter estimated against the F1-score (F-measure). **Figure 4 (a)** shows a plot of F1-score (F-measure) against the soft-margin hyper-paramerter C for SVM-L. **Figure 4 (b)** shows a plot of F1-score (F-measure) against C, but with gamma=0.01. **Figure 4 (c)** plots F1 against the number of estimator stages for GBDT. Note low F1-score (F-measure) for certain codes in all category of classifiers, such as **C3** and **C**. Despite an iterative refinement of feature space for each (classifier,code) pair and an exhaustive parameter sweep for model selection, performance for some critical codes remained low. An error analysis of codes with low F1-score (F-measure) revealed presence of co-occurring codes and code proportions to be predictive features, which necessitates adding a topic model and a meta-learner that consumes the output of the base classifiers and the topic proportions of the topic model.

can include features not seen but anticipated for a particular code, it is difficult to achieve the same with respect to topic models. However, the topic proportion θ for each document can now serve as an additional feature for a code that can be learned by a meta-learner where it looks at the collective output of all the base models prior to assigning a final set of codes to a given story.

Stacked Generalization

Stacked generalization refers to the method of training meta-classifiers to learn from the output of base classifiers towards increasing the predictive power beyond that capable by individual base classifiers. This approach has been shown to produce meta-learners with greater predictive power in many domains and applications (Chen, Wang, and Wang 2009). The error analysis of during the training of the base classifiers revealed the two significant needs given the relatively small size of the training set and the relatively low frequencies of certain labels: (1) features that took into the account the presence of co-occurring code, and (2) features that took into account code proportions in stories. We have addressed the limitations of simply using the L-LDA model

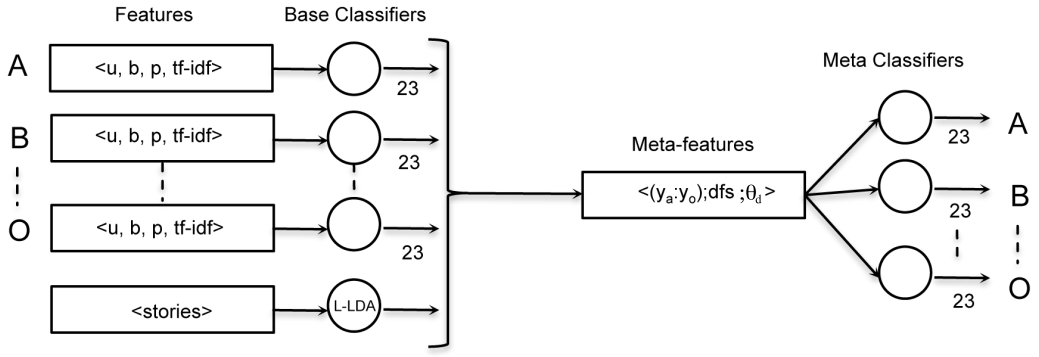


Figure 5: **Stacked generalization**: The SVM-L, SVM-R and GBDT for each code is combined into a meta-feature set that is fed into a meta-classifier. The meta-features are made of individual base classifier, Features for base classifiers include unigrams (u), bigrams, (b) part-of-speech bigrams(p) and tf-idf filtered via chi-squared feature selection and additional hand-coding of features. The output of the base classifiers are vector of predictions. $< y_A : y_O >$ and the decision function scores for each prediction. This along with the topic distribution θ_d from the L-LDA model for a given story then become meta-features for the suite of meta-learners.

in its error analysis subsection. With the fine-tuning of the base classifiers and the L-LDA model, we now have the basic machinery to train a series of meta-learners which takes as its input the predictions of the base classifiers and delivers a verdict by examining patterns of the proportion and co-occurrence of codes from the base classifiers.

Prior work in machine learning with respect to the application stacked generalization to standard datasets has shown that it is not sufficient to merely combine the predictions of base classifiers, but also their class-confidence or class-probability scores (Ting and Witten 2011). The output of the base classifiers are vector of predictions. $< y_A : y_O >$ and the decision function scores for each prediction. For the SVM-L and SVM-R models, we use the decision function defined by Vapnik (Guyon, Boser, and Vapnik 1993) as follows:

$$\text{sgn}(\sum_{i=1}^{n} y_i \alpha_i K(x_i, x) + \rho) \text{ for } y_i \in (1, -1)$$

For GBDT trees, we use the decision function defined by Ridgeway present in the scikit-learn package (Friedman 2001) This along with the topic distribution θ_d from the L-LDA model for a given story then become meta-features for the suite of meta-learners. We choose SVM-L, SVM-R and GBDT as meta-learners.The approach training the meta-learners from the metadata is exactly the same as that for the base learners. We perform a 10-fold stratified cross validation during training, while using the same splits of training, validation and test for purposes of stack-level comparison. There was no feature engineering for the meta-learners as they were fed the entire output of the base classifiers. A parameter sweep was performed for all three types of meta-learners, similar to the exercise carried out for training the base classifiers.

Results

Results from training the metaclassifiers confirms the analysis and intuition derived from the error analysis done during the training of the base classifiers with respect to co-occurrence of codes and their proportions. The best meta-learner, when compared against the best individual base learner shows a performance gain (in terms of F1 scores) for all the codes. For some codes such as **F3** and **N**, the gains are stark. Every code except **C** and **C3** had an F1 scores of above 0.66. The meta-learning exercise shows that whilst codes and sub codes (such as F2 denoting nudies taken without permission and F3 denoting forwarded without permission) might make it difficult to model their nuances using a single-level model, stacked generalization can be an effective way of modeling such nuances given the limited size of training data. In the next section we apply the results from the meta-learners to the raining 5147 uncoded instances of the dataset to derive a co-occurrence matrix 'map' of the distribution of adolescent mental health issues in this ATL dataset.

Discussion

The aforementioned work and its findings merit ***six*** aspects worth a discussion. **(1)** First, we discuss the map of adolescent issues from the co-occurrence gradient matrix from Figure 7. **(2)** We discuss adolescent relationships, and in particular how our findings clarify and expand upon the initial purpose for which ATL was established. **(3)** We underline the significance of a fine-grain analysis of the dataset and how it helped in the modeling process; **(4)** the question of why the stacked generalization framework performed better; **(5)** the statistical limitations of this work and caveats to how the findings might be interpreted, and **(6)** of how this approach might be used to power practical real-time applications. We conclude with a call for future work and plausible policy implications for school curriculum.

Code	SVM-L Base	SVM-R Base	GBDT Base	Best Meta Learner
A	0.83	0.83	0.85	**0.85**
B	0.69	0.69	0.67	**0.75**
B2	0.41	0.41	0.233	**0.9**
C	0.05	0.13	0.06	**0.36**
C1	0.49	0.49	0.45	**0.71**
C2	0.64	0.64	0.65	**0.75**
C3	0.14	0.11	0.18	**0.42**
C4	0.48	0.48	0.16	**0.72**
D	0.63	0.63	0.57	**0.77**
E	0.42	0.38	0.51	**0.72**
F	0.71	0.71	0.64	**0.91**
F2	0.26	0.26	0.28	**0.66**
F3	0.45	0.44	0.33	**0.86**
G	0.7	0.7	0.66	**0.73**
H	0.51	0.51	0.42	**0.75**
I	0.37	0.4	0.51	**0.66**
I1	0.57	0.52	0.72	**0.85**
J	0.65	0.64	0.46	**0.72**
K	0.68	0.67	0.55	**0.82**
L	0.63	0.63	0.6	**0.8**
M	0.67	0.66	0.59	**0.81**
N	0.48	0.49	0.52	**0.8**
O	0.56	0.56	0.26	**0.66**

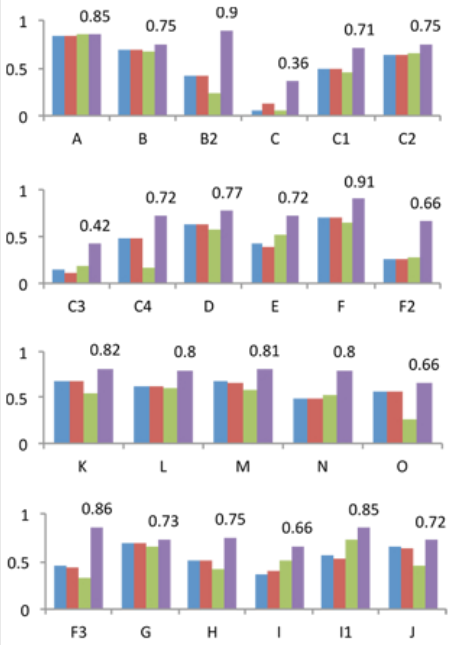


Figure 6: **Meta-learner outperform base models for all codes**: results from the meta-learning training process. The above figure shows a comparison of the best-meta learner with that of each of the individual base learners with respect to the F1 score. The meta-learner performed better than either base learner in all codes. For some codes, such as **C1**, **F3** and **N**, the gains are considerable. This confirms the analysis and intuition derived from the error analysis done during the training of the base classifiers with respect to co-occurrence of codes and their proportions.

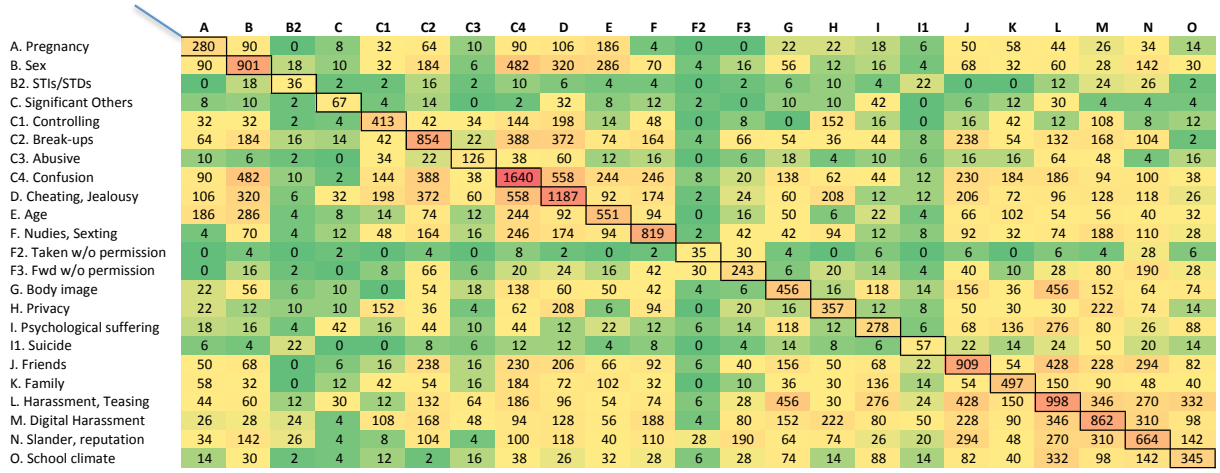
(1) A map of adolescent issues - distributions of events, themes and their possible associated emotional reactions

The set of base and meta-learners of the stack generalization framework was applied to the uncoded stories of the dataset to obtain codes for the complete dataset. A co-occurrence gradient matrix is shown in figure 7. The top 5 most frequently occurring codes are **C4** (romantic confusion), **D** *(cheating and jealousy)*, **L** *(harassment and teasing)*, **B** *(sex)* and **M** *(digital harassment)*. Not limited to these top 5 codes in the ATL, a major theme underlying them is that of social and romantic relationships if not a desire for psychological and sexual intimacy respectively.

The matrix in Figure 7 also shows a relationship between individual codes with every other code. The code **I1** *(self-injurious behavior)* co-occurs most frequently with **M** *(digital harassment)*, while **I** *(psychological pain and suffering)* co-occurs most frequently with **L** *(digital harassment and teasing)* and **K** *(involving family)*. O *(school climate)* co-occurs most frequently with **L** *(harassment and teasing)* followed by **N** *(slander, reputation)*, while **M** *(digital harassment)* is a distant fourth. The results of this analysis align with recent research (Levy et al. 2012) in suggesting that although digital bullying issues occur, they have not replaced the salience of traditional forms of offline bullying (i.e., physical and verbal) in teens lives. That **F** *(nudies)* and **C4** *(romantic confusion)* also co-occur frequently suggests that despite an increasing use of digital technologies, teenagers are not sure what is normative or appropriate for digital sharing within the teenage community. In fact, **C4** is the top co-occurring code for **B** *(sex)* and for **C2** *(breakups)*, suggesting that desires and attributes of romantic and physical intimacy in teenagers though predating the internet age, assumes new dimensions with the use of digital technologies. One important caveat (which we discuss in detail in the next section), is based on the self-selection bias involved in this participant group. This is a limitation of this work and the distribution of teenage issues depicted here cannot be generalized to the entire teenage population

(2) The overwhelming emphasis on adolescent romantic relationships: policy implications

Our findings support other large scale analyses that suggest romantic relationships and the desire for physical, social and emotional intimacy dominate teens stories about both their online and offline lives (Weinstein and Selman under review). Therefore, a pertinent question to ask is: do these



	A	B	B2	C	C1	C2	C3	C4	D	E	F	F2	F3	G	H	I	I1	J	K	L	M	N	O
A. Pregnancy	280	90	0	8	32	64	10	90	106	186	4	0	0	22	22	18	6	50	58	44	26	34	14
B. Sex	90	901	18	10	32	184	6	482	320	286	70	4	16	56	12	16	4	68	32	60	28	142	30
B2. STIs/STDs	0	18	36	2	2	16	2	10	6	4	4	0	2	6	10	4	22	0	0	12	24	26	2
C. Significant Others	8	10	2	67	4	14	0	2	32	8	12	2	0	10	10	42	0	6	12	30	4	4	4
C1. Controlling	32	32	2	4	413	42	34	144	198	14	48	0	8	0	152	16	0	16	42	12	108	8	12
C2. Break-ups	64	184	16	14	42	854	22	388	372	74	164	4	66	54	36	44	8	238	54	132	168	104	2
C3. Abusive	10	6	2	0	34	22	126	38	60	12	16	0	6	18	4	10	6	16	16	64	48	4	16
C4. Confusion	90	482	10	2	144	388	38	1640	558	244	246	8	20	138	62	44	12	230	184	186	94	100	38
D. Cheating, Jealousy	106	320	6	32	198	372	60	558	1187	92	174	2	24	60	208	12	12	206	72	96	128	118	26
E. Age	186	286	4	8	14	74	12	244	92	551	94	0	16	50	6	22	4	66	102	54	56	40	32
F. Nudies, Sexting	4	70	4	12	48	164	16	246	174	94	819	2	42	42	94	12	8	92	32	74	188	110	28
F2. Taken w/o permission	0	4	0	2	0	4	0	8	2	0	2	35	30	4	0	6	0	6	0	6	4	28	6
F3. Fwd w/o permission	0	16	2	0	8	66	6	20	24	16	42	30	243	6	20	14	4	40	10	28	80	190	28
G. Body image	22	56	6	10	0	54	18	138	60	50	42	4	6	456	16	118	14	156	36	456	152	64	74
H. Privacy	22	12	10	10	152	36	4	62	208	6	94	0	20	16	357	12	8	50	30	30	222	74	14
I. Psychological suffering	18	16	4	42	16	44	10	44	12	22	12	6	14	118	12	278	6	68	136	276	80	26	88
I1. Suicide	6	4	22	0	0	8	6	12	12	4	8	0	4	14	8	6	57	22	14	24	50	20	14
J. Friends	50	68	0	6	16	238	16	230	206	66	92	6	40	156	50	68	22	909	54	428	228	294	82
K. Family	58	32	0	12	42	54	16	184	72	102	32	0	10	36	30	136	14	54	497	150	90	48	40
L. Harassment, Teasing	44	60	12	30	12	132	64	186	96	54	74	6	28	456	30	276	24	428	150	998	346	270	332
M. Digital Harassment	26	28	24	4	108	168	48	94	128	56	188	4	80	152	222	80	50	228	90	346	862	310	98
N. Slander, reputation	34	142	26	4	8	104	4	100	118	40	110	28	190	64	74	26	20	294	48	270	310	664	142
O. School climate	14	30	2	4	12	2	16	38	26	32	28	6	28	74	14	88	14	82	40	332	98	142	345

Figure 7: **A distribution of distressing adolescent events and their emotional reactions**: The stacked generalization framework was applied to the entire ATL dataset. The above figure shows a co-occurrence matrix for the codes predicted from the stacked generalization framework.The gradient is from green (lowest score) through yellow (50th percentile) to orange (highest score). The diagonal cells denote the number of times each code appeared in the entire dataset. The co-occurrence patterns show relationships such as that between **L** (teasing and harassment) and **G** (body image), **C** (breakups) and **C4** (romantic confusion) etc. The most frequently occurring theme is the code **C4** (romantic confusion) as well as **D** (cheating, jealousy). Two key points are worth noting: **(1)**Romantic relationships was the single most implicated issue in the whole corpus (see show **C, C1,C2, C3 and C4)** co-occur frequently with other codes, and **(2) I** (psychological suffering) is implicated with self-notions of **G** (body image), problems or lack of support from **K** (family) and **L** harassment and teasing.

findings suggest needs to be researched in the online space at a policy level. While there are widespread efforts focusing on intergroup acceptance and tolerance of others through diversity programs and campaigns, the findings in this study point clearly to interpersonal and romantic relationships as salient sources of confusion and distress that have not found sufficient emphasis from an educational perspective (Lobron and Selman 2007). Recent efforts at embedding interpersonal empathy awareness into the school curriculum (Brackett et al. 2013) is a welcome step in this direction and calls for research on awareness programs, including those on digital platforms on managing and coping with romantic relationships during the teenage years. However, it should be noted as well that despite the ATL website intention to support teenagers suffering from digital abuse–or drama, our thematic coding analysis reveals 70% of the stories shared by distressed teenagers did not involve digital media.

(3) Mixed-initiative, participatory modeling

We can hardly emphasize how much the emic coding process by prevention science experts and their continued participation in the error analysis and evaluation of the predictive modeling helped in the design of the models themselves. The initial coding process was done entirely by the prevention science experts, generating thematic codes they deemed relevant given their proximity to the research surrounding teenagers' mental health in the digital age. An initial parallel but independent attempt by the machine learning specialists to produce a codebook resulted in several hiccups - for instance, should bullying at home and bullying at school be codes by themselves? The systematic process adopted by the prevention science experts to produce fine-grain codes and even subtler sub-codes to understanding the salient nuances in the dataset was highly reliable. Nevertheless, this emic thematic coding was an expensive and time-consuming process. More important, then, was the way the prevention science experts played a role in the continued error analysis at every stage of the modeling process, providing valuable insights that were parameterized into the feature space design of the models themselves.

(4) Why did stacked generalization fare better?

Stacked generalization is a proven approach to combining the predictive power of weaker base models to produce a model with higher performance. For a discriminatory model that separates two class labels in a feature space, one often provides both positive and negative features for such a discrimination to take place. But the error analysis for classifiers for the various codes and even subtler sub-codes showed patterns where the presence and proportion of a sub-code merited the presence of other codes. While features from the other sub-codes say, from **N** *(slander)* could be added to **O** *(school climate)*, those trials often resulted in further degradation of the base classifiers. While semi-supervised topic models satisfy these requirements, they are trained heavily on the vocabulary of the training sets and do not allow affordances to input features that generalize a class label beyond what is present in the training set. A meta learner can exploit co-occurring patterns of codes and their proportions by combining the output of the base ensemble (including the topic model), which was indeed our finding. During the process of coding, a story was examined from

several angles such as detection of a **victim** or **perpetrator** as well the purpose of posting such as **reporting**, **seeking help** etc. Furthermore, the presence of subtle sub-codes (for example **F2** denoting taking nudies without permission versus **F3** forwarding nudies without permission to cause harm meant that there were cues beyond just the presence of certain unigrams or bigrams that led to the assignment of a sub-code. Another reason why a meta learner might be far better than the base ensemble for such sub-codes could be that it takes into account some of these subconscious coding attributes by examining patterns of code co-occcurence and their proportions. We doubt that this is the only way to achieve the performance levels that we did and deem stacked generalization as one of many possible approaches.

(5) Self-selecting bias limitation

Much of clinical psychology and psychiatry focuses on longitudinal studies with randomized controls and variables for researching adolescent mental health issues at the individual level. On the other hand, there are studies done across a population, with controls for participant selection and variables to offset selection biases. This study is in between - it is neither a longitudinal study at the individuals' level, nor is it a study representative of a population (teenagers in the United States). These stories are from a self-selected group of teenagers who voluntarily choose to share their personal accounts online. Given the anonymity of the participants involved (age and gender are not mandatory fields to share a story), there are very limited affordances to offset at least some of the self-selection bias. For example we suspect that girls are far greater users of ATL, but we also do not know the gender (or age) of over 37% of our sample. We take the view that there ought to be other large scale analyses such as this for other self-help apps and websites for teenagers under duress for such a generalization to take place. Another interesting aspect of self-selection for internet-based self-help is that there is very little scientific literature (Donkin et al. 2012) that looks deeply at statistical self-selecting bias for internet users and how that might be death with.

(6) Possible uses in practical applications

The combined thematic emic coding and stacked generalization learning approach provides interesting opportunities towards practical, real-time systems. A multifaceted understanding of the emic themes, can either allow us to (1) more effectively connect the person with resources and (2) present them with a similar stories to help them feel that they are not alone in their plight. There are help forums and websites (see related work) that have emerged in recent years in an assistive role for not just teenagers, but for a range of populations and issues from people with autism to sufferers of acne and other disorders. Given the volume of data points provided such participants online, there are opportunities to analyze teenage angst on an even larger scale with temporal dimensions, looking at, for example, the seasonality and prevalence of teenage mental issues and their evolution over time. Furthermore, such an approach can be used to create reflective user interfaces on these websites (Dinakar et al. 2012a; 2012b) and apps for treating mental disorders online.

Summary

We adopt an interdisciplinary, mixed-initiative approach to analyzing issues of teenage distress. We analyze 7147 personal stories shared by distressed teenagers on a popular teen-help website aimed at supporting teenagers experiencing digital harassment and other issues of distress. A team of prevention science psychology experts analyze the dataset qualitatively, adopting an emic approach, thematic coding to produce a codebook with the most important issues they deem relevant from the dataset for a sample of 2000 stories from the dataset. A suite of base binary classification models are trained on the coded stories with 10-fold stratified sampling, exhaustive parameter sweeps for model selection and iterative feature space engineering. An error analysis of the base binary classifiers merits the training of a semi-supervised labeled LDA model to account for code occurrence and their proportions. The output of this topic model and the entire ensemble of base classifiers is fed into training a ensemble of meta-learners to predict codes for a given story. An evaluation shows that this stacked generalization learning outperforms the level-1 base ensemble for all the codes. This stacked generalization framework is then used to generate codes for the entire collection of 7147 personal stories. Our methods suggest that teenagers are talking a lot on this platform about romantic relationships and intimacy clearly this is something thats on their minds.There isnt a lot of formal support or curricula around these issues in schools, which might explain why they are turning to the Internet looking for an outlet/source of support. We can do better and need more effective ways to support teens in their intimate and psychosocial development. While a component of this is definitely digital, that is only one piece; we also need to introduce digital supports without overlooking the traditional offline challenges that continue to impact teenagers (Weinstein and Selman under review).

Acknowledgements

We wish to thank MTV for their help in obtaining the ATL dataset. We thank also the entire class from Prevention Science and Practice at Harvard University.

References

Bogdanova, D.; Rosso, P.; and Solorio, T. 2012. Modelling fixated discourse in chats with cyberpedophiles. In *Proceedings of the Workshop on Computational Approaches to Deception Detection*, 86–90. Association for Computational Linguistics.

Boyatzis, R. E. 1998. *Transforming qualitative information: Thematic analysis and code development*. Sage.

Boyd, D., and Marwick, A. 2011. Social privacy in networked publics: Teens attitudes, practices, and strategies. *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*.

Brackett, M. A.; Bertoli, M.; Elbertson, N.; Bausseron, E.; Castillo, R.; and Salovey, P. 2013. Emotional intelligence. *Handbook of Cognition and Emotion* 365.

Chen, J.; Wang, C.; and Wang, R. 2009. Using stacked generalization to combine svms in magnitude and shape feature spaces for classification of hyperspectral data. *Geoscience and Remote Sensing, IEEE Transactions on* 47(7):2193–2205.

Choudhury, M. D.; Gamon, M.; Counts, S.; and Horvitz, E. 2013. Predicting depression via social media. In *International AAAI Conference on Weblogs and Social Media*.

Coie, J. D.; Watt, N. F.; West, S. G.; Hawkins, J. D.; Asarnow, J. R.; Markman, H. J.; Ramey, S. L.; Shure, M. B.; and Long, B. 1993. The science of prevention: a conceptual framework and some directions for a national research program. *American Psychologist* 48(10):1013.

Dinakar, K.; Jones, B.; Havasi, C.; Lieberman, H.; and Picard, R. 2012a. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Trans. Interact. Intell. Syst.* 2(3):18:1–18:30.

Dinakar, K.; Jones, B.; Lieberman, H.; Picard, R. W.; Rosé, C. P.; Thoman, M.; and Reichart, R. 2012b. You too?! mixed-initiative lda story matching to help teens in distress. In *International Conference on Weblogs and Social Media*.

Dinakar, K.; Reichart, R.; and Lieberman, H. 2011. Modeling the detection of textual cyberbullying. In *International AAAI Conference on Weblogs and Social Media, Social Web Workshop*.

Doherty, G.; Coyle, D.; and Sharry, J. 2012. Engagement with online mental health interventions: an exploratory clinical study of a treatment for depression. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, 1421–1430. ACM.

Donkin, L.; Hickie, I.; Christensen, H.; Naismith, S.; Neal, B.; Cockayne, N.; and Glozier, N. 2012. Sampling bias in an internet treatment trial for depression. *Translational psychiatry* 2(10):e174.

Farzan, R.; Kraut, R.; Pal, A.; and Konstan, J. 2012. Socializing volunteers in an online community: a field experiment. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, 325–334. ACM.

Friedman, J. H. 2001. Greedy function approximation: a gradient boosting machine. *Annals of Statistics* 1189–1232.

Gross, R., and Acquisti, A. 2005. Information revelation and privacy in online social networks. In *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*, 71–80. ACM.

Guan, J., and Huck, J. 2012. Children in the digital age: exploring issues of cybersecurity. In *Proceedings of the 2012 iConference*, 506–507. ACM.

Guyon, I.; Boser, B.; and Vapnik, V. 1993. Automatic capacity tuning of very large vc-dimension classifiers. *Advances in neural information processing systems* 147–147.

Levy, N.; Cortesi, S.; Crowley, E.; Beaton, M.; Casey, J.; and Nolan, C. 2012. Bullying in a networked era: A literature review. *Berkman Center Research Publication* (2012-17).

Lobron, A., and Selman, R. 2007. The interdependence of social awareness and literacy instruction. *The Reading Teacher* 60(6):528–537.

Madden, M.; Lenhart, A.; Cortesi, S.; Gasser, U.; Duggan, M.; Smith, A.; and Beaton, M. 2013. Teens, social media, and privacy. *Pew Research Center*.

Mrazek, P. B., and Haggerty, R. J. 1994. *Reducing risks for mental disorders: Frontiers for preventive intervention research: Summary*. National Academies Press.

Neff, K. D., and McGehee, P. 2010. Self-compassion and psychological resilience among adolescents and young adults. *Self and identity* 9(3):225–240.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Ramage, D.; Hall, D.; Nallapati, R.; and Manning, C. D. 2009. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, 248–256. Association for Computational Linguistics.

Ramage, D.; Dumais, S. T.; and Liebling, D. J. 2010. Characterizing microblogs with topic models. In *ICWSM*.

Selman, R. L., and Dray, A. J. 2006. Risk and prevention. *Handbook of child psychology*.

Selman, R. L., and Schultz, L. H. 1998. *Making a friend in youth: Developmental theory and pair therapy*. Aldine de Gruyter.

Selman, R. L. 1980. *The growth of interpersonal understanding: Developmental and clinical analyses*. Academic Press New York.

Suzuki, L. K., and Calzo, J. P. 2004. The search for peer advice in cyberspace: An examination of online teen bulletin boards about health and sexuality. *Journal of Applied Developmental Psychology* 25(6):685–698.

Ting, K. M., and Witten, I. H. 2011. Issues in stacked generalization. *arXiv preprint arXiv:1105.5466*.

Toutanova, K.; Klein, D.; Manning, C. D.; and Singer, Y. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, 173–180. Association for Computational Linguistics.

Viera, A. J.; Garrett, J. M.; et al. 2005. Understanding interobserver agreement: the kappa statistic. *Fam Med* 37(5):360–363.

Webb, M.; Burns, J.; and Collin, P. 2008. Providing online support for young people with mental health difficulties: challenges and opportunities explored. *Early intervention in psychiatry* 2(2):108–113.

Weinstein, E. C., and Selman, R. L. under review. under review. *under review*.