You Too?! Mixed-Initiative LDA Story Matching to Help Teens in Distress

Karthik Dinakar[+] Birago Jones[+] Henry Lieberman[+] Rosalind Picard[+] Carolyn Rose[*] Matthew Thoman[▾] Roi Reichart[+]

[+]Massachusetts Institute of Technology [*]Carnegie Mellon University [▾]Northeastern University

Abstract

Adolescent cyber-bullying on social networks is a phenomenon that has received widespread attention. Recent work by sociologists has examined this phenomenon under the larger context of teenage drama and it's manifestations on social networks. Tackling cyber-bullying involves two key components – automatic detection of possible cases, and interaction strategies that encourage reflection and emotional support. Key is showing distressed teenagers that they are not alone in their plight. Conventional topic spotting and document classification into labels like "dating" or "sports" are not enough to effectively match stories for this task. In this work, we examine a corpus of 5500 stories from distressed teenagers from a major youth social network. We combine Latent Dirichlet Allocation and human interpretation of its output using principles from sociolinguistics to extract high-level themes in the stories and use them to match new stories to similar ones. A user evaluation of the story matching shows that theme-based retrieval does a better job of finding relevant and effective stories for this application than conventional approaches.

Introduction

Studies have shown that cyber-victimization and cyber-bullying on social networks involving adolescents are strongly associated with psychiatric and psychosomatic problems. A cyber-bully status has been shown to be associated with hyperactivity, conduct problems, low pro-social behavior, frequent smoking and drunkenness, headache, and not feeling safe at school, while a cyber-victim status has been shown to include emotional and peer problems, headache, recurrent abdominal pain, sleeping difficulties. [1] To understand what constitutes adolescent cyber-bullying and cyber-victimization, we draw a distinction between how adults view cyber-bullying versus how teenagers perceive it. The latest research by social scientists has shown what might be perceived as cyber-bullying by adults is often viewed as mere 'drama' by teenagers. [2] Psychiatrists have espoused the need for the induction of strategies to foster cognitive empathy to deal with cyber-bullies as well as cyber-victims. [3]

We synthesize the aforementioned ideas from sociologists and psychiatrists to frame a hypothesis that the tackling of cyber-bullying involves two key components – detection and an in-context reflective user-interaction strategy to encourage empathy on social networks. In this paper, we use a corpus of real-world stories by distressed teenagers from MTV for two purposes – to understand the distribution of 'drama-like' themes from the collection of stories as well as how that can be used for theme-based story matching that can power a reflective user-interface.

We use probabilistic latent Dirichlet allocation and principles from sociolinguistics to draw dominant themes from the corpus and use that to match a new story to a similar story – with a view of trying to show distressed teenagers that they are not alone in their plight. A user-evaluation shows positive results for the themes extracted and for matching a given story with similar stories.

Related Work

Because of the complexity and the inter-disciplinary angles to the problem of tackling cyberbullying, it is worth examining related work from three broad perspectives, namely the social sciences, psychiatry and computer science (computational linguistics and user-interaction). Social scientists and psychiatrists have been studying the phenomenon of cyberbullying from two angles – gauging the extent of its prevalence and studying the cognitive and psychiatric detrimental effects it has on adolescents. There have been many attempts at devising mitigation strategies at the school and parental level, although evaluation of these strategies has proved to be difficult to ascertain [4].

In the field of computational linguistics, related work for the detection of textual cyberbullying has involved the use of statistical supervised machine learning topic classification to detect sensitive topics such as sexuality, race and culture, socio-normative conflicts, physical appearance and intelligence [5]. The Genesis project has investigated the understanding of automatic plot

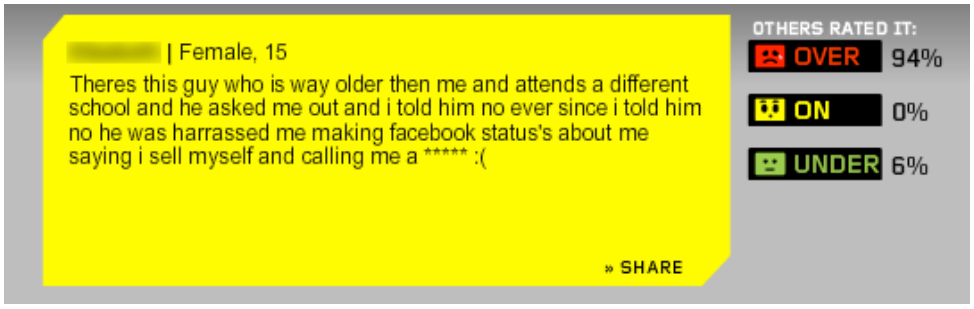

Figure 1 A sample story from a distressed teenager on athinline.org describing her negative experience. The reactions from others who viewed this story are shown on the right. 94% of the people who voted think who voted think that this story was 'over the line', while 6% think that this was 'under the line' and not very serious.

construction of stories through commonsense reasoning [6], although not for addressing the problem of cyberbullying.

Current strategies by popular social networking websites involve key-word spotting to automatically flag threads that smack of cyberbullying and providing affordances to report inappropriate content. Yet, having millions of users on their networks means that even the flagged instances typically run into the thousands everyday, thereby making the problem of prioritizing serious instances extremely difficult. In the next section, we discuss in detail the parameters derived from sociologist and psychiatric research on cyberbullying and how that fits into our algorithmic approach of identifying themes and using them to match online discourse with similar stories.

Teenage drama and its thematic distributions

The traditional Olweus definition of bullying highlights three aspects of bullying, namely the unwanted and negative aggressive behavior of a bully, its repeating nature, and power imbalances that makes self-defense difficult for a victim. If one were to detect textual cyberbullying under this definition, it becomes essential to detect the underlying topic of discussion that might cause a power imbalance. Previous work in this realm has espoused that topics that are immutable with respect to individual characteristics, such as race, culture, sexuality and physical appearance etc, are helpful to identify power imbalances [7].

But recent research suggests a much broader definition of bullying beyond that of Olweus that has interesting implications for any algorithmic detection of cyberbullying. Drawing the distinction that teenagers like to refer to adult-labeled bullying as mere drama rather than bullying broadens the scope of what constitutes cyberbullying. Researchers have shown that drama includes a distribution of aggressive and passive-aggressive behaviors, ranging from posting what teens often refer to as "inappropriate" videos and photos and the resulting fallout; conflicts that escalate into public standoffs; cries for attention; relationship breakups, makeup's and jealousies etc [2].

This suggests that for effective detection of textual cyberbullying on social networks one needs to factor in the distribution of these teenage drama themes, beyond classification techniques to label an interaction on a social networking website into just a single label. In the next section, we discuss an approach to extract these themes from a corpus of real-world stories by distressed teenagers from the popular MTV website athinline.org. We begin by describing the sociolinguistic characteristics of the corpus and then describe our approach of extracting themes.

The MTV teenage stories corpus

The popular youth culture network MTV's website, athinline.org allows distressed adolescents and young adults in distress to share their stories anonymously with a view of getting crowd-sourced feedback and advice. When a teenager posts a story on their website, anyone can read it and vote on its severity in three ways – over the line (severe), on the line (moderate to mild) and under the line (not very serious). Although the site first started as an attempt to help teenagers from digital harassment such as *sexting* and social network bullying, the range of topics in the stories that teenager talk about on the site have a much broader scope, from dating to very serious cases of physical abuse. The age of those posting stories on the site ranges from 12-24, although more than half of it comes from teenagers.

We analyze a completely anonymized corpus of 5500 personal stories from this website, with data on the votes

that each story received. Upon an initial examination of the corpus, most stories contained a set of themes. For example, in Figure 1, the two dominant themes for the story could be imagined as a combination of 'high school drama' and 'bullying on social networks'.

Diglossia & hedging

A further examination of some of the stories in the corpus revealed interesting sociolinguistic attributes characterizing teenage discourse. Consider the following two stories from the corpus:

"I had this one guy trick me into thinking he was "considering" liking me over the summer. I have big boobs and he really wanted to see them online. It was all done over sexting and all that crap. I didn't do it but he was pressuring me."

*"i have this guy friend that is reallly nice but hes always asking me to send him sexy pic or pics of my **** or **** n i say no be cause im not a **** so sorry n the next day he doesnt talk to me n calls me a **** not cool right??????????????????"*

Both of the stories have 'Sending / uploading nude / naked pictures of boyfriend or girlfriend' as their dominant theme though the styles employed by the former is more formal than the later. In the former, there is no mention of the word 'picture', but the story still has the same theme. The two stories highlight diglossia in online teenage discourse [10], a situation where a single community employs two dialects or styles, which in this case happens to be girls.

Similarly, consider the following story by a teenage girl whose severity is rather grave:

"my bf doesn't luv me anymore. im kinda sucks, kinda scared want me life 2 end"

This story is representative of hedging [11], or softening one's expression of something acerbic. The two dominant theme in this story appear to be 'breakup heartache' and 'feeling scared'. The example tells us that the absence of keywords such as suicide is not representative of how grave it really is. In fact, there were at least 14 other stories by girls with the same combination of themes that were indicative of suicidal tendencies that were voted as 'over the line' by third-party individuals. These kinds of intuitions can serve as a yardstick for prioritizing targeted help for the most severe stories. In the next section, we describe an approach to extract themes from these stories keeping in mind the aforementioned sociolinguistic themes in the corpus.

Extracting high-level themes

To extract high-level themes from the MTV corpus, there are two plausible approaches. The first was to subject the corpus to human annotation by requiring annotators to mark each story with the themes present in it. Whilst this would curate the corpus for supervised hierarchical classification, it would require all the 5500 stories to be annotated by the same set of people for the best results. Instead, we choose a different two-step approach: a) to apply latent Dirichlet Allocation (LDA) to get a set of word clusters and b) interpret these sets word clusters to assign a theme to each individual word cluster using principles from sociolinguistics. The rationale behind choosing LDA is to avoid manual annotation of every story and to assign theme distributions to each individual story.

We use the following standard LDA model for this purpose as follows [8]:

Let T denote the number of themes in teenage stories, and D denote the number of stories in your database, and N the total number set of words in the corpus. Let P(z) denote the distribution over themes z in a particular story, and P(w|z) for the probability mass over word w given theme z. We can then write the following distribution of words within a given story:

$$P(w_a) = \sum_{b=1}^{T} P(w_a \mid z_a = b) P(z_a = b)$$

If we define parameters φ and θ as

$$\theta^{(d)} = P(z) \quad \phi^{(b)} = P(w \mid z = b)$$

We follow the model by Griffiths and Steyvers in drawing a Dirichlet priors Dir(β) and Dir(α) for both φ and θ. We use the Gibbs sampling method [9] and use α = 50/T and β = 0.01 for each iteration of our approach, which we describe below:

Assigning themes to word clusters

We begin the process of extracting themes by setting the number of topics **T** in increments of 10. The process of determining whether a satisfactory number of themes have emerged from the LDA model is as follows:

Step 1: Begin with number of topics T = 10, with hyper-parameter updating every 10 iterations

Step 2: Assign themes to each of the T word clusters by examining the words under each cluster

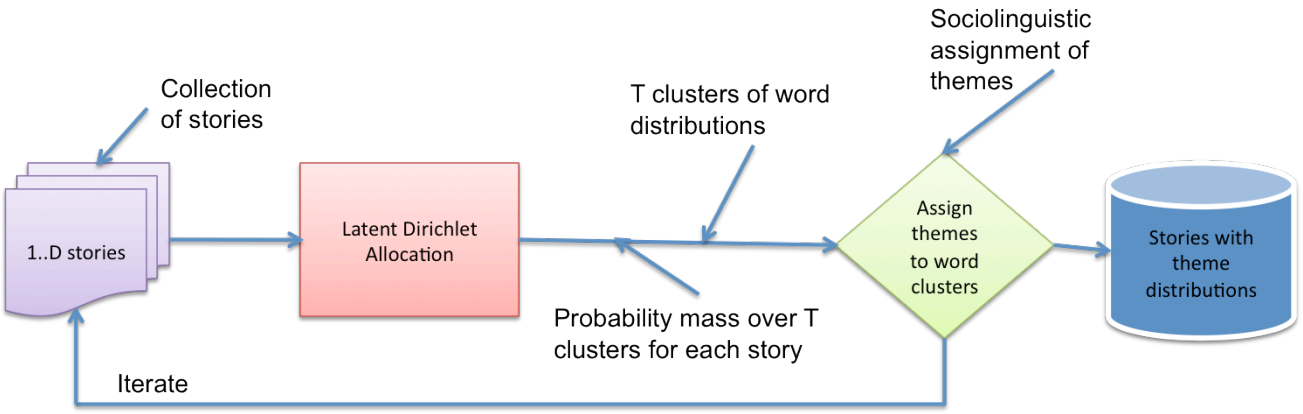

Figure 2 Extracting high-level themes from the MTV corpus. The stories are first run through LDA to get clusters of word distribution. Using principles from sociolinguistics, the clusters are interpreted into themes. This process is repeated until the a satisfactory number of themes can be extracted consistent with observations from the latest social science research on teenage drama

Step 3: Repeat until a satisfactory number of well-defined themes have emerged by updating the T by 10

Since LDA is an inverted generative process that assumes a distribution of topics structure in the way humans write text, the process of assigning themes to word clusters can happen only if the latent space is semantically meaningful and interpretable in the context of addressing teenage drama. Hence, we asked a group of 3 people to examine each cluster of words and use a subset of them to create sentence(s) that denotes the semantic theme that they thought was best for the cluster. The participants were encouraged to use as few other words apart from the cluster of words the purpose of creating sentence(s) from each of the clusters to denote a theme.

For example, consider the following word cluster the merged from the LDA model:

bf trust ive cheated times months break yrs past issues caught cheat hasnt bi fone multiple numbers forgive touch

For the aforementioned cluster of words, the sentences constructed were as follows:

I caught my boyfriend cheating on me.

I found her phone and noticed she had a lot of messages from numbers on her phone. They were from boys in my school.

After the above exercise, each of the participants was asked to assign a theme for the sentences. For the example, the theme was 'cheating and trust issues'. This loop of interpreting the LDA results was done until the number of topics T reached 30.

Evaluation of the assignment

The kappa values (Cohen's kappa) for theme assignment was > 0.6 for 25 out of the 30 topics generated. Of the remaining topics, clustered shorthand notations such as 'lol, idk, rofl' etc and four clusters did not yield anything of semantic value.

Story theme distributions

Based on the LDA output, each story gets a probability mass for each of the 30 themes. We make this into a visualization to show the distribution of the themes present in the corpus. The most prevalent theme present in the corpus was 'uploading of naked or nude pictures by boyfriend / girlfriend', with advice sought on 'duration of relationships' followed by 'bullying on social media' and bullying connected with appearance'. This is well documented in the social science literature investigating the problem of bullying, especially the fact that most bullying with regard to physical appearance tends to emanate from girls bullying girls. [12].

Story theme distributions

Based on the LDA output, each story gets a probability mass for each of the 30 themes. We make this into a visualization to show the distribution of the themes present in the corpus. The most prevalent theme present in the corpus was 'uploading of naked or nude pictures by boyfriend / girlfriend', with advice sought on 'duration of relationships' followed by 'bullying on social media' and bullying connected with appearance'. This is well

Theme	% of stories	#	Theme	% of stories
Using naked pictures of girlfriend or boyfriend	7.6%	14	Hookups	3.8%
Duration of a relationship/dating	5.0%	16	Shorthand notations (not a theme)	3.8%
Bullying on social media	4.9%	17	Age & dating	3.7%
Bullying connected with appearance	4.9%	18	One-sided relationships	3.5%
High school & college drama	4.7%	19	Communication gaps & misunderstandings	3.4%
Bullying on email & cell-phone	4.6%	20	Hanging out with friends	3.3%
Involving parents, siblings or spouses	4.4%	21	Jealousy	3.2%
Feeling scared, threatened or worried	4.3%	22	Talking negatively behind one's back	3.2%
Sexual acts, pregnancy	4.1%	23	Anguish & depression	3.1%
Inability to express how you feel	4.0%	24	Ending a relationship	3.1%
First dates & emotions	3.9%	25	Long-term relationships under duress	3.0%
Falling in love	3.8%	26	Post-breakup issues	2.9%
Cheating & trust issues	3.8%			

Table 1. Table showing the distribution of themes (theme that gets the most probability mass in a story is the most dominant theme for that story) for the 5500 stories in the MTV corpus

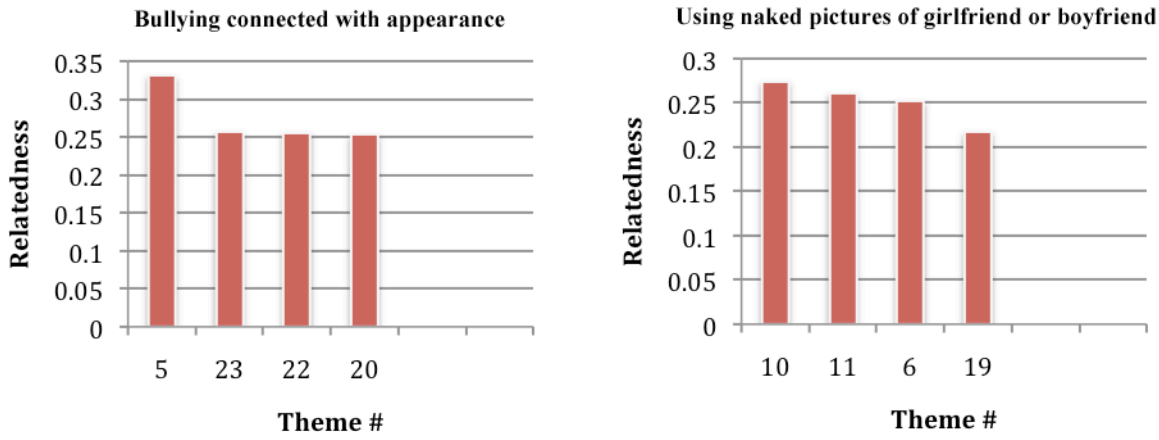

Figure 3 - Top 4 themes that are most related with a given theme. The relatedness is calculated by how many times another theme co-occurred with the current theme amongst the top 3 themes for all stories. That count was divided by the total number of times the given theme occurred in the top 3 themes for all stories.

documented in the social science literature investigating the problem of bullying, especially the fact that most bullying with regard to physical appearance tends to emanate from girls bullying girls. [12].

Co-occurring themes

As previously explained, it becomes important to understand a story from a multiple thematic perspective. We investigate the top 3 themes that each story gets, to see what other themes co-occur with it. By investigating the co-occurrence of a theme in a story, a moderator or the interface could use it for targeted help. For example, most stories with the co-occurring themes 'Breakups, anguish and depression' and 'Feeling scared, threatened and worried' are by girls. By connecting the meta-data about the individual's gender with the co-occurrence of themes, moderators can point the individual for advice specifically tailored for girls dealing with depression connected with breakups, which might be better than generalized advice on depression.

The relatedness score to calculate co-occurring stories across the corpus was done as follows for each theme A: let $\mathbf{C_{(B,A)}}$ be the count of how many times another theme **B** co-occurs with A amongst the top 3 themes for all stories.

"okay so i had this boyfriend, and i kissed another guy, but i told him. and we broke up. then months later we got back together and his friend came up to me and kissed me. and then her told my boyfriend, so he broke up with me. but i want him back!"

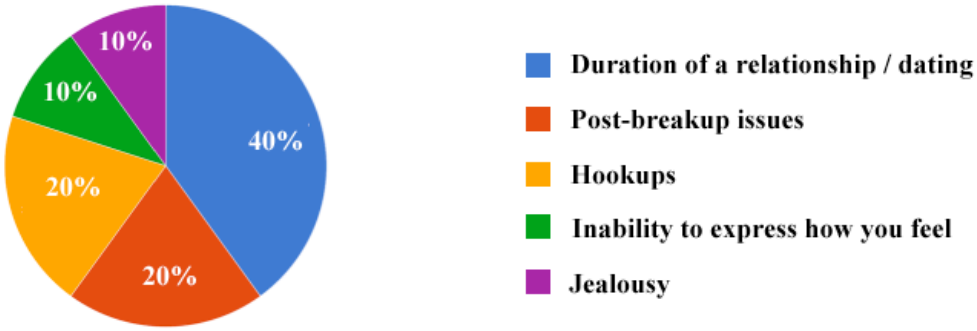

Figure 4 - Thematic breakdown of a story whose dominant themes include duration of a relationship and post-breakup issues the top 2 dominant themes. As one can see, the most accurate themes get assigned the most probability mass by the LDA model

Let $\mathbf{C_{(A)}}$ denote the total number of times A occurs in the top 3 themes for all stories. Then the relatedness $\mathbf{R_{(B,A)}}$ is defined by:

$$R_{(B,A)} = \frac{C_{(B,A)}}{C_{(A)}}$$

Figure 3 shows the top 5 themes that were most related to the themes 'Breakups, anguish and depression' and 'Uploading naked pictures by boyfriend / girlfriend' respectively. As seen in the picture, bullying connected with appearance is the most related topic connected with 'Breakups, anguish and depression'. Since the overwhelming gender talking about breakups in an anguished way were girls, we can began to connect research from psychiatry that associates appearance as a very important aspect of female adolescent psyche [12]. For the theme of 'uploading naked pictures of boyfriend or girlfriend', a related theme is 'first dates and emotions' and 'bullying via email and cell phones'. We wish to state that we are not validating any stereotypes here, but merely connecting the dots from disparate fields of study for the problem of cyberbullying that might lead to more targeted help for those teenagers that need help the most.

Thematic breakdown of a story

Another useful tool to understand each story better would be to visualize the thematic breakdown of each story as given by the LDA. We build an interface that allows aggregate views of all the themes, drill-down into a particular theme, as well as view details of a single story. Such a breakdown would be useful to a moderator to not only analyze a given story, but also allow visualizations to drill-down into particular subsets of themes, moving seamlessly between views to better understand the dynamics of the stories and to set moderation policies based on current dynamics. For example, stories of teenage girls that have the first two dominant themes as 'breakups, anguish and depression' and 'feeling scared, threatened or worried' may merit further scrutiny given the serious nature of previous stories with the same dominant themes.

Theme-based story matching

Research from adolescent psychiatry suggests that distressed teenagers can be helped with targeted, in-context and specific advice about their situation [3]. When a distressed teenager comes to a help site with his or her story, it would be apt to show the individual a similar story encountered by another teenager with the same experiences. What this calls for is story matching based on a distribution of themes rather than by singular labels such as 'positive' or 'negative'. By matching stories based on themes, it can induce a level of reflection on the part of the distressed individual that can cause both self-awareness and a feeling that one is not alone in their plight. Knowing that there are other teenagers who have experienced similar ordeals can go a long way in helping the individual deal with distress.

Let $\boldsymbol{S}$ denote the set of existing stories. We use the following approach to match new stories to existing ones using the following approach:

Step 1: Apply the LDA model to a new story $\boldsymbol{A}$ to get a probability distribution $\boldsymbol{x}$ for $\boldsymbol{T}$ themes.

Step 2: Select a subset of previously submitted stories $\boldsymbol{B} \subset \boldsymbol{S}$ where the top 5 themes $\boldsymbol{T_1}$ through $\boldsymbol{T_5}$ match the top 5 themes of the new story in the same order.

	% Strongly Agree		% Agree		% Strongly Disagree		% Disagree		% Neither agree nor disagree	
	LDA+	Control	LDA+	Control	LDA+	Control	LDA+	Control	LDA+	Control
Q1	45.0%	0%	22.1%	3%	15.9%	46%	17%	51%	0%	0%
Q2	35.3%	0%	23.0%	8.3%	15%	31.0%	13.2%	35.1%	13.5%	25.6%

Table 2 – Evaluation table showing the user-study responses to questions Q1 ('The themes of the story presented matched the themes of the story I wrote') and Q2 ('After reading the presented story, I can imagine that someone in a similar situation would not feel alone').

New Story	Matched story
I like wearing makeup. I dont want to be teased for my weight. Some girls hurt me that no one will like me at all. I don't want to be called names about my looks!	Growing up I havent been the skinnest girls. When I first started Middle School people would always make fun of my wieght. Because of that I would always tell my self that I was fat and ugly and i really took it to heart. And it still follows me.

Table 3 – An example from our user-study of a new story and it's best match as measured by asymmetric KL-divergence. Bullying associated with physical appearance is overwhelmingly female in the corpus with 'High school & college drama' as strongly related topic.

Step 3: Let ***y*** denote the probability distribution for each story in the above subset ***B.*** The degree of similarity of the new story to each story in subset B is gauged by using the *Kullback-Liebler* asymmetric divergence metric as follows:

$$KL(x,y) = \sum_{i=1}^{T} x_i \log \frac{x_i}{y_i}$$

Step 4: Sort the stories in ***B*** in ascending order of their KL divergence. The story with the least KL divergence is most likely to be similar to the new story ***A***.

Evaluation

We design an evaluation protocol to test the two crucial aspects of this work – a) the effectiveness of our algorithmic approach of story matching in relation to a conventional technique's baseline, and b) the effectiveness of showing the matched stories to help reflective thinking in distressed teenagers to feel that they're not alone in their plight. We also provide a qualitative evaluation of the usefulness of identifying the distribution of themes by two community moderators of two popular teenage community websites respectively.

Control: We compare our algorithmic approach of story matching against *cosine* similarity of *tf-idf* vectors for the new and old stories as control.

Participant selection: We selected a total of 12 participants of which 8 were female. 5 of them were teenagers at the time of the study. 2 of the participants were teachers at a local public school, while 3 were graduate students working with young children. 2 of the participants were graduate students researching machine learning and human-computer interaction respectively.

User-study protocol: The participants were each subjected to an online user-study as follows. First, each participant was asked to view the *MTV AThinline* website to familiarize themselves with a) the kind of stories and b) the type of linguistic styles employed by teenagers on the site. Second, each participant was asked to enter a new story using a 250 world limit (as in the case of *Athinline*) on any relevant themes as they deemed appropriate. 5 matched stories of which 3 were retrieved using our algorithmic approach and 2 using the control approach were shown to the participant after each new story.

Each matched story was evaluated with respect to two questions **Q1**: if the retrieved stories matched the story entered by participants in having similar themes and **Q2**: if they could imagine a distressed teenager feeling a little better if they were shown the matched stories – that they were not alone in their plight. Each participant was asked to write a minimum of 3 new stories, thereby evaluating a total of 15 matches. A total of 38 new stories were entered, with a total 190 matches.

Results & Discussion

Results show strong results for story matching using our LDA and KL-divergence approach of extracting themes and matching new stories to old ones. Our approach fared

better for both **Q1** and **Q2** against a simple cosine similarity using tf-idf vectors for the story matching.

An error analysis showed that new stories for which the matches were rated as 'Strongly Agree' had very clear themes with linguistic styles very similar to the stories in the corpus. Those new stories for which the matched stories that received a 'Strongly disagree or 'Disagree' vote did not have clear themes or used a vocabulary that wasn't common in the corpus. For example, consider the following new story:

'I wanted to be in the school dance team, but I was not accepted. I think its cause the captains don't like my bf.'

The above story is an example of a story that didn't have a clear theme relative to those extracted from the corpus. This best match for this story was *'a guy at my school let me flurt with him and he knew i liked him and know he wont even talk to me at all cuz we have no classes together is he over the line or not'* . Though the algorithm did produce a coarse level match with respect to school and liking a male, the story still received a 'Strongly disagree' vote. This calls for a deeper level of reasoning for fine-grain story matching.

At the end of the user-study, participants were positive about an overall feedback, with comments such as 'Even when it missed some things, I could see why it was trying to go that way' and 'It was really neat stuff'.Two moderators of two teenage community-based social networks viewed the distribution of themes from the corpus extremely positively. The moderators likened the distribution of themes, even on a 'flagged' set of instances from their websites, as key to understanding the dynamics of negative behavior on their websites.

Future Work

We concede that the best evaluation of this work would be to test it on a live on the MTV AThinline website, with in-situ distressed teenagers using the interface without knowing the story matching that is happening in the background. We are currently in the process of deploying our work on the site for a real-world, in-situ evaluation and remain excited about the practical nature of the research questions that this work has produced.

Acknowledgements

We wish to thank all the people involved in this research project, the White House, Department of Education and MTV for their continued focus and support.

References

1. Andre Sourander, Anat Brunstein Klomek, Maria Ikonen, Jarna Lindroos, Terhi Luntamo, Merja Koskelainen, Terja Ristkari, and Hans Helenius Psychosocial Risk Factors Associated With Cyberbullying Among Adolescents: A Population-Based Study Arch Gen Psychiatry, Jul 2010; 67: 720 – 728.
2. Marwick, Alice E. and boyd, danah, The Drama! Teen Conflict, Gossip, and Bullying in Networked Publics (September 12, 2011). A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society, September 2011.
3. Ang, R. P., & Goh, D. H. (2010). Cyberbullying among adolescents: The role of affective and cognitive empathy, and gender. Child Psychiatry and Human Development, 41(4), 387-397.
4. Patchin, J. W. & Hinduja, S. (2012). Cyberbullying Prevention and Response: Expert Perspectives. New York: Routledge (ISBN: 978-0415892377).
5. Dinakar, K.; Reichart, R.; Lieberman, H.. Modeling the Detection of Textual Cyberbullying. International AAAI Conference on Weblogs and Social Media, North America, July. 2011.
6. Capen. W.H., Story Understanding in Genesis: Exploring Automatic Plot Construction through Commonsense Reasoning, Masters Thesis, MIT EECS 2011.
7. Janis Wolak, Kimberly J. Mitchell, David Finkelhor, Does Online Harassment Constitute Bullying? An Exploration of Online Harassment by Known Peers and Online-Only Contacts, Journal of Adolescent Health, Volume 41, Issue 6, Supplement, December 2007, Pages S51-S58, ISSN 1054-139X, 10.1016/j.jadohealth.2007.08.019.
8. Mark Steyvers, Tom Griffiths. Probabilistic Topic Models. In Landauer, T., Mcnamara, D., Dennis, S., Kintsch, W., Latent Semantic Analysis: A Road to Meaning. (2006).
9. Ian Porteous, David Newman, Alexander Ihler, Arthur Asuncion, Padhraic Smyth, and Max Welling. 2008. Fast collapsed gibbs sampling for latent dirichlet allocation. In Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '08). ACM, New York, NY, USA.
10. Schiffman, H. (1997). Diglossia as a sociolinguistic situation. In The Handbook of Sociolinguistics, F. Coulmas (ed.), 205–216. Oxford: Blackwell.
11. Wardhaugh, R., An Introduction to Sociolinguistics, Blackwell Textbooks in Linguistics, 2009.
12. Tiggemann, M., & Miller, J. (2010). The Internet and adolescent girls' weight satisfaction and drive for thinness. Sex Roles, 63, 79-90.