

MIT Open Access Articles

*Commonsense Reasoning for Detection,
Prevention, and Mitigation of Cyberbullying*

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

Citation: Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying. ACM Trans. Interact. Intell. Syst. 2, 3, Article 18 (September 2012), 30 pages.

As Published: <http://dx.doi.org/10.1145/2362394.2362400>

Publisher: Association for Computing Machinery (ACM)

Persistent URL: <http://hdl.handle.net/1721.1/79894>

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Terms of use: Creative Commons Attribution-Noncommercial-Share Alike 3.0



Commonsense Reasoning for Detection, Prevention and Mitigation of Cyberbullying

KARTHIK DINAKAR, BIRAGO JONES, CATHERINE HAVASI, HENRY LIEBERMAN, ROSALIND PICARD MIT Media Lab

Cyberbullying (harassment on social networks) is widely recognized as a serious social problem, especially for adolescents. It is as much a threat to the viability of online social networks for youth today as spam once was to email in the early days of the Internet. Current work to tackle this problem has involved social and psychological studies on its prevalence as well as its negative effects on adolescents. While true solutions rest on teaching youth to have healthy personal relationships, few have considered innovative design of social network software as a tool for mitigating this problem. Mitigating cyberbullying involves two key components – robust techniques for effective detection, and reflective user interfaces that encourage users to reflect upon their behavior and their choices.

Spam filters have been successful by applying statistical approaches like Bayesian networks and hidden Markov models. They can, like Google’s *GMail*, aggregate human spam judgments because spam is sent nearly identically to many people. Bullying is more personalized, varied, and contextual. In this work, we present an approach for bullying detection based on state-of-the-art NLP and a Commonsense knowledge base, which permits recognition over a broad spectrum of topics in everyday life. We analyze a more narrow range of particular subject matter associated with Bullying (e.g. appearance, intelligence racial and ethnic slurs, social acceptance and rejection), and construct *BullySpace*, a commonsense knowledge base that encodes particular knowledge about bullying situations. We then perform joint reasoning with Commonsense knowledge about a wide range of everyday life topics. We analyze messages using our novel *AnalogySpace* Commonsense reasoning technique. We also take into account social network analysis and other factors. We evaluate the model on real-world instances that have been reported by users on *Formspring*, a social networking website that is popular with teenagers.

On the intervention side, we explore a set of reflective user-interaction paradigms with the goal of promoting empathy amongst social network participants. We propose an “air traffic control”-like dashboard, which alerts moderators to large-scale outbreaks that appear to be escalating or spreading, and to help them prioritize the current deluge of user complaints. For potential victims, we provide educational material that informs them about how to cope with the situation, and connects them with emotional support from others. A user evaluation shows that in-context, targeted and dynamic help during cyberbullying situations fosters end-user reflection that promotes better coping strategies.

Categories and Subject Descriptors: **H.1.2 [Models and Principles]:** User/Machine Systems—*Human Information Processing*; **1.5.1 [Pattern Recognition]:** Models

General Terms: Artificial Intelligence

Additional Key Words and Phrases: Commonsense reasoning, affective computing, AI,

ACM Reference Format:

Dinakar, K., Havasi, C., Jones, B., Lieberman, H., Picard, R. 2011 Commonsense Reasoning for Detection, Prevention and Mitigation of Cyberbullying

DOI=10.1145/0000000.0000000 <http://doi.acm.org/10.1145/0000000.0000000>

1. THE PROBLEM OF CYBERBULLYING

Cyberbullying has grown as a major problem in recent years, afflicting children and young adults. Recent surveys on the prevalence of cyberbullying have shown that almost 43% of teens in the United States alone were subjected to cyberbullying at some point [1]. The adverse impact that cyberbullying has on victims, especially adolescents in their formative years, is well documented [2]. The American Academy of Child and Adolescent Psychiatry says that targets of cyberbullying often deal with significant emotional and psychological suffering [3]. Studies have shown that cyber-victimization and cyber-bullying on social networks involving adolescents are strongly associated with psychiatric and psychosomatic disorders [4]. Children whose coping mechanisms may not be as strong as that of an

Permission to make digital or hardcopies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credits permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.

©2010 ACM 1544-3558/2010/05-ART1 \$10.00

DOI=10.1145/0000000.0000000 <http://doi.acm.org/10.1145/0000000.0000000>

adult sometimes suffer grievously, sometimes resulting in tragic outcomes like self-injurious behavior and even suicides.

According to the National Crime Prevention Council, cyberbullying can be defined as the following: “when the Internet, cell phones or other devices are used to send or post text or images intended to hurt or embarrass another person” [5]. Social scientists such as danah boyd [6] have described four aspects of the web that change the very dynamics of bullying and magnify it to new levels: persistence, searchability, replicability and invisible audiences.

Cyberbullying is a more persistent version of traditional forms of bullying, extending beyond the physical confines of a school, sports field or workplace, with the victim often experiencing no respite from it. Cyberbullying gives a bully the power to embarrass or hurt a victim before an entire community online, especially in the realms of social networking websites.

1.1 Current efforts in Cyberbullying

Previous work addressing cyberbullying has centered on extensive surveys unearthing the scope of the problem and on its psychological effects on victims. At a recent White House Conference on Bullying Prevention [7], US President Barack Obama and First Lady Michelle Obama expressed their deep concern about this problem. The US government has created a resource bank [8] that provides a survey of the current works of psychologists and educators concerning bullying, and provides resources for children, parents, teachers, and law enforcement.

Little attention if any, to date, has been devoted to technical solutions in social network software to automatically detect bullying or provide interventions directly in software interaction [71]. Many social networks have an “Online Safety Page” that leads to resources such as the anti-bullying sites of the government or other organizations. They deal with the bullying problem primarily by people responding to explicit user complaints. Even so, in popular networks, the volume of complaints received daily quickly overwhelms the ability of small groups of complaint handlers to deal with them. The few automated detection facilities are extremely simple and ineffective, often using regular-expressions to catch a list of profane words. Solutions originally developed for spam filtering are repurposed for bullying detection [9], with little precision and many false positives. In the detection section below, we explore the application of conventional statistical machine learning techniques for detection of bullying.

On the intervention side, little has been done directly in interactive software. Many schools provide explicit educational material to educate students about the problem and provide advice. In some US school districts such education is mandated, and often takes the form of school assemblies. Typical of such educational efforts is Jay Banks’s STAMP [10] which offers a set of reasonable, but vague guidelines: such as ‘*stay away from bullies*’, ‘*telling someone about a negative bullying experience*’, ‘*avoiding bad situations*’, ‘*making friends*’ and so on.

One of the most effective ways to provide education is through stories, and among current sites, one of the best examples is MTV’s The Thin Line [11]. This site solicits stories of possible bullying from youth, and encourages engaging discussion about whether a particular situation is “Over the Line” of acceptability. This “crowd-sourced ethics” approach is good because it respects the value of opinions of the children themselves, does not preach or assume there is a single “right” answer, and offers actionable advice. But the site discusses a vast variety of situations that may or may not be relevant to a particular individual or problem, and the site itself is separated from the social networks where the actual interaction takes place.

In summary, our criticism of current efforts to deal with the Cyberbullying problem is that detection efforts are largely absent or extremely naïve. Intervention efforts are largely offline, and fail to provide specific, actionable assessment and advice.

After we provide technical descriptions of our specific efforts, in the Related Work section below we will discuss specific work in Natural Language Understanding, Interactive Education, and other relevant areas to particular aspects of our work.

2. RELATED WORK

Much of the work related to cyberbullying as a phenomenon is in the realm of social sciences and psychology. As such, this problem has not been attacked from the perspective of statistical models for detection and intervention involving reflective user interaction. Related academic work to tackle cyberbullying must be viewed from three perspectives – ethnographic studies by social scientists to gauge its prevalence, a psychological analysis of its negative impacts, and related tangential work in the NLP and the user interaction community.

a. Social Sciences & Psychiatry

A lot of research in the social sciences has been devoted to understanding the causes of cyberbullying and the extent of its prevalence, especially for children and young adults [12]. Research in psychiatry has explored the consequences, both short and long term, that cyberbullying has on adolescents and school children and ways of addressing it for parents, educators and mental health workers [13]. Such studies, which often involve extensive surveys and interviews, give important pointers to the scope of the problem and in designing awareness campaigns and information toolkits to schools and parents, as well as offering important algorithmic insights to parameterize detection models to catch candidate instances of cyberbullying.

b. Text categorization tasks & parameterization of online interaction analysis

Machine learning approaches for automated text categorization into predefined labels has witnessed a surge both in terms of applications as well as the methods themselves. Recent machine learning literature has established support-vector machines as one of the most robust methods for text categorization, used widely for email spam filters. The European Union sponsored project PRINCIP has used support vector machines using a bag-of-words approach to classify web pages containing racist text [14]. Indeed, support vector machine was one of our better performing methods for recognizing one of three categories of bullying remarks [15].

Recent work in the NLP community for classification tasks involving online interaction analysis such as identifying fake reviews has shown the effectiveness and importance of drawing intuitive parameters from related domains such as psycholinguistics [16]. In this work, we rely heavily on observations and intuitions from related work in the social sciences and psychology for both problem decomposition as well as feature space design.

c. Similar real-world applications

Apart from spam filters, applications that are of a similar nature to this work are in automatic email spam detection and automated ways of detecting fraud and vandalism in Wikipedia [17].

Very few applications have attempted to address the Bullying problem directly with software-based solutions. The FearNot project [18] has explored the use of virtual learning environments to teach 8-12 year old children coping strategies for bullying based on synthetic characters. This uses interactive storytelling with animated on-screen characters, where the user gets to play one of the participants in the bullying scenario. The user may select any one of a number of response strategies to a bullying challenge, e.g. fight back, run away, tell a teacher, etc. Though it provides the user with participatory education about the situations, the situations are artificially constructed. They are not part of the users' real lives. It does not make any attempt to analyze or intervene in naturally occurring situations where serious injury might be imminent and might be prevented.

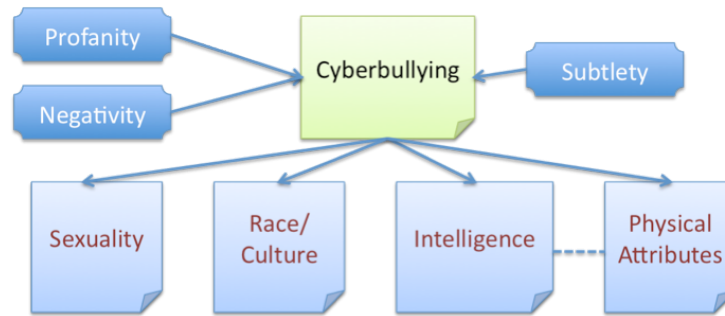


Figure 1 Problem decomposition. A given textual comment or a post that is part of a discourse interaction on social networking websites is likely candidate for cyberbullying if the underlying topic is of a sensitive nature and its has one or more contributing features of profanity, negativity and subtlety.

3. DETECTING BULLYING

Cyberbullying is generally defined as the repeated, injurious use of harassing language to insult or attack another individual [19]. Cyberbullying includes uploading of obscene pictures; unethical hacking into another individual’s personal account on a social network, impersonation and verbal harassment. We focus specifically on the problem of textual cyberbullying, in the form of verbal harassment. Unlike spam, which is generic in the sense that it is multicast to hundreds of millions of people, bullying is aimed at a particular individual, much less often at group of individuals. We decompose cyberbullying by topics. Because of the targeting of a specific individual, the topics involved in bullying are those of a sensitive nature that is personal to the victim. Social scientists talk extensively about the usage of sensitive topics to establish a power differential between the bully and victim [23]. The set of such personal topics includes race and culture, sexuality, physical appearance, intelligence and social rejection.

We break down the problem of topic modeling into detecting high-level contributing features, namely profanity and contextually relevant patterns of abuse, the use of negative language devoid of profanity, as well as the employment of subtlety designed to insult another person. We experiment with statistical supervised machine learning methods to detect bullying and describe their limitations in terms of finding insulting language when there is no explicit profane or negative language. We then describe how a model of commonsense reasoning can address this limitation.

This section of the paper begins with a description of the problem, and the corpora used for this work. Following that is a treatment of the statistical machine learning techniques used to classify text into one of the aforementioned topics. We describe how conventional machine learning techniques can detect contextually relevant patterns of abuse and the use of negative language devoid of profanity, but fail to address instances of bullying that are subtle and which need commonsense reasoning for detection. We then proceed to give an overview of the Open Mind Commonsense knowledge base as well the *AnalogySpace* inference technique, before describing a commonsense reasoning model to address the difficult problem of detecting subtlety used to insult another individual.

3.1 Rationale behind problem decomposition

When a comment or a message tends to involve sensitive topics that may be personal to an individual or a specific group of people, it deserves further scrutiny. In addition, if the same comment also has a negative connotation and contains profanity, the combination of rudeness and talking of sensitive, personal topics can be extremely hurtful. Equally potent if not more, are comments or posts that are implicitly inappropriate – lacking in profanity or negativity but designed to mock or to insult.

For most children in middle school and young adults, psychological research espouses that the sensitive list of topics often assume one of the following: physical appearance, sexuality, race &

culture and intelligence [20]. Repeated posting of such messages can lead to the victim internalizing what the bully is saying, which can be harmful to the well-being of the victim.

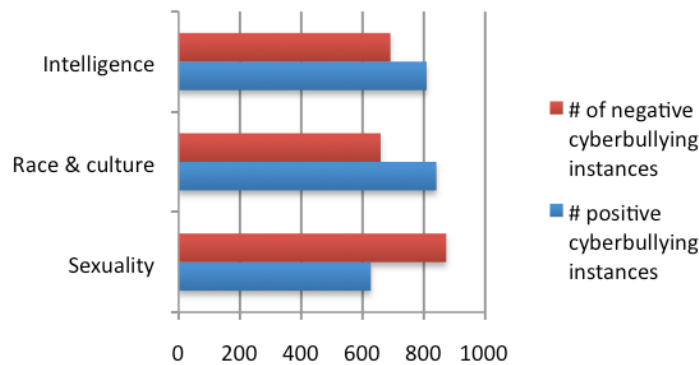
3.2 Corpora

We use two datasets for this work: YouTube and Formspring. The YouTube dataset for experiments with statistical machine learning was obtained by scraping the social networking site www.youtube.com for comments posted on controversial (videos discussing sensitive issues such as race, culture, same-sex marriage, role of women in society etc) and relatively non-controversial videos (linear algebra and photoshop tutorials etc). Though YouTube gives the owner of a video the right to remove offensive comments from his or her video, a big chunk of viewer comments on YouTube are not moderated. Videos on controversial topics are often a rich source for objectionable and rude comments.

The comments downloaded from all the videos were arranged in random order prior to annotation. Three annotators of whom one was an educator who works with middle school children, annotated each comment along the lines of three labels defined as follows:

- a. **Sexuality** - Negative comments involving attacks on sexual minorities and sexist attacks on women.
- b. **Race and culture** - Attacks bordering on racial minorities (e.g. African-American, Hispanic and Asian) and cultures (e.g. Jewish, Catholic and Asian traditions) and stereotypical mocking of cultural traditions.
- c. **Intelligence** - Comments attacking the intelligence and mental capacities of an individual.

After annotation, 1500 comments under each category for which the inter-rater kappa agreement was 0.4 or higher were selected for the purpose of the use of supervised learning methods.



Label/Annotation	# of positive cyberbullying instances	# of negative cyberbullying instances
Sexuality	627	873
Race & Culture	841	659
Intelligence	809	691

Figure 2 – The YouTube were annotated and grouped into three categories of 1500 instances each under sexuality, race & culture and intelligence. 627, 841 and 809 instances were found to be positive for sexuality, race & culture and intelligence respectively.

An effective strategy towards computational detection of cyberbullying must address both the explicit and direct forms of abuse, as well as the subtler, indirect ways of insulting an individual. Although

the YouTube corpus is an excellent source to find direct forms of abuse involving objectionable and profane content, it lacks the personalized discourse interaction seen on a more community oriented social networking website. However, the YouTube corpus does serve as an excellent source for training supervised learning models to detect explicit forms of verbal abuse. The annotation task was to examine each comment, assign a label (bullying or no bullying) . In the case of the a positive label (bullying), the annotators were asked to assign the topic for the comment (sexuality, race & culture or intelligence).

The MIT Media Lab partnered with the social networking website Formspring, whose popularity amongst teenagers and young adults has grown by leaps and bounds since their launch in November of 2009. From Formspring, we received a dataset of anonymized instances of bullying that were either user-flagged or caught by their moderation team. The Formspring dataset contained instances of bullying that were more targeted and specific than the YouTube corpus. It also had numerous instances of bullying involving subtlety, with use of stereotypes and social constructs to implicitly insult or malign the target.

A ranked feature list derived from the use of supervised learning methods on the YouTube dataset was used to filter out comments from the Formspring corpus that contained blatant occurrences of profanity and negativity to obtain instances that were implicit in their intentions to bully. The same team of three annotators was asked to pick instances from this filtered dataset that pertained to topics of sexuality, namely topics involving lesbian, gay, bisexual & transgender (LGBT) stereotypes. The Formspring dataset contained instances that were already flagged as inappropriate by their users, allowing us to use their labels for whether an instance was cyberbullying or not.

We choose to focus the detection on the topics and stereotypes related to the LGBT community because firstly, bullying and cyberbullying of young adolescents based on LGBT stereotypes has been well documented both in the psychology community [21] and secondly, the usage of LGBT stereotypes without profanity highlights the need to move beyond conventional statistical learning methods for effective detection [22]. We do not endorse any of the stereotypes through this work, and seek only to use them for detection of ways of accusing or speculating about the sexuality of another individual. In the next section, we explore the use of the use of statistical supervised machine learning techniques.

3.3 Statistical Machine Learning techniques

Interaction analysis on social networks is a complex phenomenon to model mathematically. The field of sociolinguistics that has been studying interaction analysis argues vigorously that the use of language between individuals in a social setting is parameterized by a rich set of characteristics, including identity, ascription to a particular community, personality and affect. They argue that it is specificity and uniqueness that matter the most for effective interaction analysis. But machine learning techniques for language are often reductionist approaches that place a heavy emphasis on abstraction, generalization and stable patterns in the data. Finding a balance between these paradigms is crucial for analyzing discourse on social networks, highlighting the importance of effective feature space design. Indeed, recent work in the computational discourse analysis community has seen the incorporation of principles from sociolinguistics for analyzing discourse [24].

Our approach towards using statistical supervised machine learning is to show its strengths and weaknesses in detecting cyberbullying. Since explicit verbal abuse involves the use of stereotypical slang and profanity as recurring patterns, those aspects lend themselves nicely to supervised learning algorithms. We also hypothesize that instances of cyberbullying where the abuse is more indirect, and which does not involve the use of profanity or stereotypical words are likely to be misclassified by supervised learning methods.

We adopt a bag-of-words supervised machine learning classification approach to identifying the sensitive theme for a given comment. We divide the YouTube corpus into 50% training, 30% validation and 20% test data. We choose three types of supervised learning algorithms in addition to Naïve

Bayes from toolkit WEKA [25], a rule-based learner, a tree-based learner and support-vector machines with default parameters, described briefly as follows:

a) Repeated Incremental Pruning to Produce Error Reduction, more commonly known as JRip, is a propositional rule learner proposed by Cohen et.al [26]. It is a two-step process to incrementally learn rules (grow and prune) and then optimize them. This algorithm constructs a set of rules to cover all the positive instances in the dataset (those with labels {sexuality, race & culture, intelligence}) and has shown to perform efficiently on large, noisy datasets for the purpose of text classification [27].

b) J48 is a popular decision tree based classifier based on the C4.5 method proposed by Ross Quinlan [28]. It uses the property of information gain or entropy to build and split nodes of the decision tree to best represent the training data and the feature vector. Despite its high temporal complexity, J48’s performance for classifying text has shown to produce good results [29].

c) Support-vector machines (SVM) [30] are a class of powerful methods for classification tasks, involving the construction of hyper-planes that at the largest distance to the nearest training points. Several papers reference support-vector machines as the state of the art method for text classification [31,32,33]. We use a nonlinear poly-2 kernel [34] to train our classifiers, as preliminary experiments with a linear kernel did not yield statistically significant differences with a poly-2 kernel, which has also been a finding in some recent empirical evaluation of SVM kernels [35].

In the first experiment, binary classifiers using the above were trained on each of the three datasets for each of the three labels: sexuality, race & culture, and intelligence to predict if a given instance belonged to a label or not. In the second experiment, the three datasets were combined to form a new dataset for the purpose of training a multiclass classifier using the aforementioned methods. The feature space was built in an iterative manner, using data from the validation set in increments of 50 instances to avoid the common pitfall of over-fitting.

Once used, the instances from the validation set were discarded and not used again to ensure as little over-fitting as possible. The trained models were washed over data from the test set for an evaluation. The kappa statistic, a measure of the reliability of a classifier, which takes into account agreement of a result by chance, was used to gauge the performance of the methods. 10-fold cross validation was applied for training, validation and testing for both the experiments.

3.4 Feature Space design

The feature space design for the two experiments can be categorized into two kinds: general features that are common for all three labels and specific features for the detection of each label.

Feature	Type
TF-IDF	General
Ortony lexicon for negative affect	General
List of profane words	General
POS bigrams: JJ_DT, PRP_VBP, VB_PRP	General
Topic specific unigrams and bigrams	Label specific

Figure 3 - Feature Design. General features were common across all the datasets for both experiments. Label specific features consisted of unigrams that were observed in the training data.

The intuition behind this is as follows: negativity and profanity appear across many instances of cyberbullying, irrespective of the subject or label that can be assigned to an instance. Specific features can then be used to predict the label or the subject (sexuality, race & culture and intelligence).

a. General features - The general features consist of TF-IDF (term frequency, inverse-document frequency) weighted unigrams, the Ortony lexicon of words denoting negative connotation, a list of profane words and frequently occurring stereotypical words for each label.

b. TF-IDF - The TF-IDF (term frequency times inverse document frequency) is a measure of the importance of a word in a document within a collection of documents, thereby taking into account the frequency of occurrence of a word in the entire corpus as a whole and within each document.

c. Ortony Lexicon for negative affect -The Ortony lexicon [36] (containing a list of words in English that denotes the affect) was stripped of the positive words, thereby building a list of words denoting a negative connotation. The intuition behind adding this lexicon as unigrams into the feature set is that not every rude comment necessarily contains profanity and personal topics involving negativity are equally potent in terms of being hurtful.

d. Part-of-speech tags - Part-of-speech tags for bigrams, namely, PRP_VBP, JJ_DT and VB_PRP were added to detect commonly occurring bigram pairs in the training data for positive examples, such ‘you are’, ‘..... yourself’ and so on.

e. Label Specific Features -For each label, label specific unigrams and bigrams were added into the feature space that was commonly observed in the training data. The label specific unigrams and bigrams include frequently used forms of verbal abuse as well as widely used stereotypical utterances For example, the words ‘fruity’ and ‘queer’ are two unigram features for the label sexuality, because of their use for hurtful abuse of LGBT individuals.

We discuss an evaluation of the aforementioned supervised learning methods in section 5. Our hypothesis is that supervised learning methods generally fare well when it comes to detecting explicit forms of verbal abuse owing to the presence of stable patterns. We anticipate in our error analysis that instances of cyberbullying that are indirect and which do not involve the use of explicit language, of which there aren’t enough training samples, are likely to be misclassified by the models. In the next section we discuss the need for using commonsense knowledge reasoning to detect instances of cyberbullying that could not be caught using the aforementioned conventional supervised learning methods.

3.5 The Open Mind Commonsense knowledge base

When we reason about the world, we are using our knowledge of what is expected, to react to and anticipate situations. As discussed before, traditional supervised learning techniques tend to rely on explicit word associations that are present in text, but using common sense can help provide information --- about people’s goals and emotions and object’s properties and relations --- that can help disambiguate and contextualize language.

The goal of the Open Mind Common Sense (OMCS) [37] project is to provide intuition to AI systems and applications by giving them access to a broad collection of basic knowledge, along with the computational tools to work with it. This knowledge helps applications to understand the way objects relate to each other in the world, people’s goals when they go about their daily lives, and the emotional content of events or situations. OMCS has been collecting common sense statements from volunteers on the Internet since 1999. At the time of this research, we have collected tens of millions of pieces of English language common sense data from crowd sourcing, integrating other resources, and the Semantic Web.

This knowledge allows us to understand hidden meaning implied by comments, and to realize when others are making comments designed to make us feel like our behavior is outside of the "normal". When we communicate with each other, we rely on our background knowledge to understand the meanings in conversation. This follows from the maxim of pragmatics that people avoid stating information that the speaker considers obvious to the listener [38].

Common sense allows us a window to what the average person about a concept or topic. This allows us to look for stereotypical knowledge, especially about sexuality and gender roles. OMCS knows that a girl is capable of doing housework, holding puppies, wearing bows in their hair, and babysitting and that a boy is capable of crying wolf, bagging leaves, wrestling, playing video games, and shouting loudly. More direct clues can be found in the gender associations of certain words: For example, OMCS associates dresses and cosmetics more strongly with girls. We emphasize that it is not our intention to validate or approve of any of these stereotypes, but only to use such stereotypical assertions for detection of subtle, indirect forms of verbal abuse.

For the knowledge we collect to become computationally useful, it has to be transformed from natural language into more structured forms that emphasize the contextual connections between different concepts. ConceptNet represents the information in the OMCS corpus as a directed graph [39]. The nodes of this graph are concepts, and its labeled edges are assertions of common sense that connect two concepts.

Concepts represent aspects of the world as people would talk about them in natural language, and they specifically correspond to normalized forms of selected constituents of common sense statements entered in natural language. This research uses ConceptNet 4, in which one of twenty-one different relations connects two concepts, forming an assertion. Each assertion has a notation of whether or not the relationship is considered to be negated (polarity), and a score representing the public's general opinion on whether the predicate is true or not. For example, the assertion "*A skirt is a form of female attire*" connects the "*skirt*" and "*form of female attire*" nodes with the "*IsA*" relation.

ConceptNet can also be represented as a matrix where the rows are concepts in the graph. The columns represent graph "*features*" or combinations of relation edges and target concepts. Features can be thought of as properties that the object might have such as 'made of metal' or 'used for flying'. This network of concepts, connected by one of about twenty relations such as "*IsA*", "*PartOf*", or "*UsedFor*", are labeled as expressing positive or negative information using a polarity flag. The relations are based on the most common types of knowledge entered into the OMCS database, both through free text entry and semi-structured entry. For the assertion "*A beard is part of a male's face*", for instance, the two concepts are "beard" and "male", the relation is "*IsA*", and the polarity is positive. For the assertion "*People don't want to be hurt*", the concepts are "*person*" and "*hurt*", the relation is "*Desires*", and the polarity is negative.

Each concept can then be associated with a vector in the space of possible features. The values of this vector are positive for features that produce an assertion of positive polarity when combined with that concept, negative for features that produce an assertion of negative polarity, and zero when nothing is known about the assertion formed by combining that concept with that assertion. As an example, the feature vector for "*blouse*" could have +1 in the position for "*is part of a female attire*", +1 for "*is worn by girls*", and +1 for "*is worn by women*". These vectors together form a matrix whose rows are concepts, whose columns are features, and whose values indicate truth values of assertions. The degree of similarity between two concepts, then, is the dot product between their rows in the concept/feature matrix. This representation is discussed in detail by Speer & Havasi et.al [40].

3.6 The AnalogySpace inference technique

In order to reason over this data set, we needed to develop an algorithm that was both noise resistant and which took advantage of patterns inherent in how we see the world. When we determine if an object is animate, for example, we may look at the properties of that object. Does it move on its own? Is it fuzzy? Or made of metal? Is it a common pet? We also think about what objects are most similar

to it. Does it look like a rabbit? Or a robot? Is it a concrete object like a pony or an immaterial quantity such as happiness?

Each question you might ask about a concept can be thought of as a “dimension” of a concept space. Then, answering a question such as where does an object lie along the “animate vs. inanimate” dimension, can be thought of as reducing the dimensions of the space from every question you might ask, to just the question of interest; that is, projecting the concept onto that one dimension. We therefore use mathematical methods for dimensionality reduction, such as singular value decomposition (SVD) [40] to reduce the dimensionality of the concept-feature matrix. This determines the principal components, or axes, which contain the salient aspects of the knowledge, and which can be used to organize it in a multi-dimensional vector space. The resulting space can be used to determine the semantic similarity using linear operations over vectors representing concepts in the semantic space. Concepts close together in the space are treated as similar; these are also more likely to combine to form a valid inference.

Let us call the matrix whose rows are concepts, whose columns are features, and whose values indicate truth values of assertions as A . This matrix A can be factored into an orthonormal matrix U , a diagonal matrix Σ , and an orthonormal matrix V^T , so that $A = U\Sigma V^T$. The singular values in are ordered from largest to smallest, where the larger values correspond to the vectors in U and V that are more significant components of the initial A matrix. We discard all but the first k components — the principal components of A — resulting in the smaller matrices U_k , Σ_k , and V_k^T . The components that are discarded represent relatively small variations in the data, and the principal components form a good approximation to the original data. This truncated SVD represents the approximation $A_k = U_k\Sigma_k V_k^T$. As AnalogySpace is an orthogonal transformation of the original concept and feature spaces, dot products in AnalogySpace approximate dot products in the original spaces. This fact can be used to compute similarity between concepts or between features in AnalogySpace.

3.7 The Blending knowledge combination technique

While it is useful to use common sense to acquire more common sense, we benefit more when we use these techniques to learn from multiple data sets. Blending [41] is a technique that performs inference over multiple sources of data simultaneously by taking advantage of the overlap between them. Two matrices are combined using a blending factor and then a SVD is taken over both data sets. Blending can be used to incorporate other kinds of information, such as information about stereotypes, into a common sense matrix to create a space more suited for a particular application.

We can use this technique to create a specific knowledge base to collect knowledge about different types of stereotypes beyond those in the OMCS database. Blending balances the sizes and composition of the knowledge bases such that the small size of such a knowledge base is not overpowered by the (much) larger ConceptNet. Additionally, information about implicit stereotypes may bring out other lightly stereotyped knowledge in the database and allows us to expand the reach of entered stereotypical knowledge. For example, adding OMCS allows us to discover that mascara, not just makeup, is usually associated with girls in the context of fashion.

Common sense can be used to fill in the gaps in other knowledge sources, both structured and unstructured, or can be designed to cover knowledge surrounding a narrow special topic. For example, in his work with SenticNet, Erik Cambria [42] created a specialized knowledge base with information about emotions. That database has been combined with common sense and domain specific texts to create a system that understands affect in free text [43].

In the following sections, we build a knowledge base to perform commonsense reasoning over a specific slice of cyberbullying, namely that concerning gay and lesbian issues.

3.6 The BullySpace knowledge base

A key ingredient to tackling implicit ways of insulting another person is to transform commonly used stereotypes and social constructs into a knowledge representation. For example consider the following instance from the Formspring corpus:

put on a wig and lipstick and be who you really are

In the above instance, a bully is trying to speculate about or malign the sexuality of a straight, male individual implicitly, by trying to attribute characteristics of the opposite sex. (Of course, in the context of a conversation between openly gay people, such a comment may be completely innocuous.) The underlying social construct here is that, in a default heterosexual context, people don't like to be attributed with characteristics of the opposite sex. This attribution is made using the common stereotype that wigs and lipstick are for women or for men who want to dress as women.

In this work, we observe the Formspring dataset and build a knowledge base about commonly used stereotypes used to bully individuals based on their sexuality. The representation of this knowledge is in the form of an assertion, connecting two concepts with one of the twenty kinds of relations in ConceptNet. For the above example, the assertions added were as follows:

lipstick **is** used by girls
lipstick **is** part of makeup
makeup **is** used by girls
a wig **is** used by girls
a toupee **is** used by men

We build a set of more than two hundred assertions based on stereotypes derived from the LGBT related instances in the Formspring database. We emphasize that our aim is not to endorse any of these stereotypes, but merely to detect their use in bullying. We then convert these assertions into a sparse matrix representation of concepts versus relations, in the same manner as ConceptNet. We then use AnalogySpace's joint inference technique, *blending*, to merge them together to create a space that is more suited for the purpose of detecting implicit insults concerning LGBT issues. While blending, we give double post-weight to the matrix generated from the set of assertions specifically designed to capture LGBT stereotypes. Once the two matrices have been merged, we then perform an AnalogySpace inference by performing an SVD to reduce the dimensionality of the matrix by selecting only the top $k=100$ set of principal components. We now have the basic machinery required to perform commonsense reasoning.

3.7 Cosine similarity of extracted and canonical concepts

A given comment is first subjected to an NLP module to perform the standard normalization operations: removing stop-words, and tokenizing the text to have a clear separation of words from punctuation marks. Next, we extract a list of concepts from the normalized text that is also present in the concept axes of the dense matrix derived after performing the SVD, as explained in the previous section.

The next task is to choose a set of canonical concepts for comparison with the concepts that have been extracted from the comment. We select four canonical concepts, namely the affective valences *positive* and *negative*, as well as gender, namely *male* and *female*. The idea here is to compare each extracted concept for similarity with each of the canonical concepts. This is achieved by performing a dot product over the extracted concept with a canonical concept. After this comparison, we normalize the values derived for each of the canonical concepts to get an overall measure of how similar the given comment is to each of the canonical concepts.

For example, consider the comment *'Did you go shopping yesterday?'*. This comment is subjected to the process described above to yield similarity scores for the canonical concepts of good, bad, boy and girl.

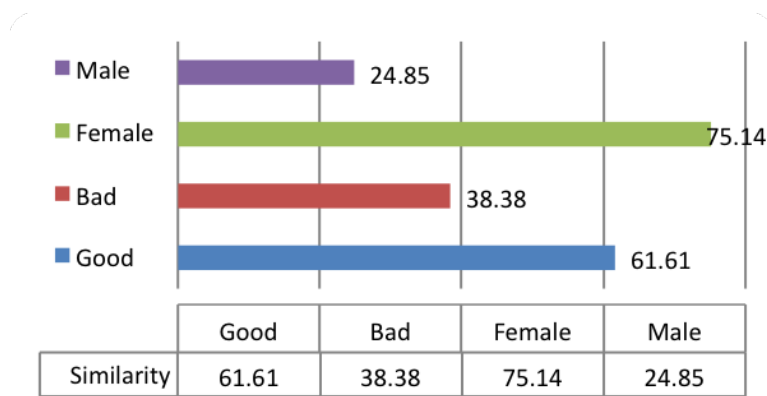


Figure 4 – Results for the comment ‘Did you go shopping yesterday?’. Shopping as a concept is more related to females. It is also more considered a good activity, owing to the fact that more users in the ConceptNet database regard it as a good activity rather than bad.

Similarity scores derived from the above examples show that shopping as a concept is more oriented towards girls than boys, and is largely considered as an enjoyable activity rather than a bad one. Based on these similarity scores, it can be inferred that this is a fairly innocent comment.

Consider another comment, ‘Hey Brandon, you look gorgeous today. What beauty salon did you visit?’. Although this contains no profanity, it does appear on the face of it, to be a comment more attributable to a girl than a boy.

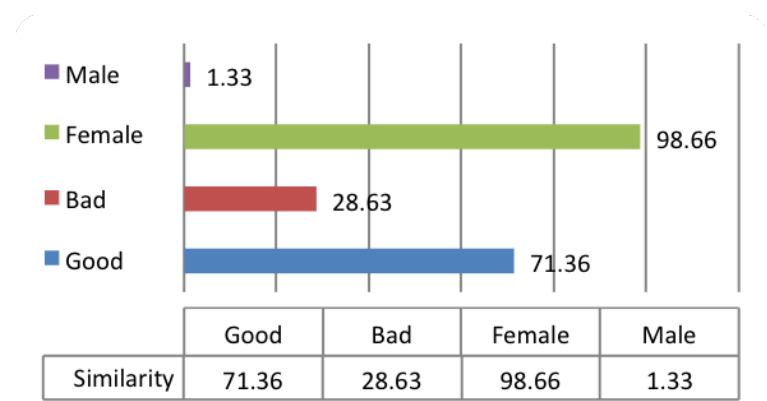


Figure 5 – Results for the comment ‘Hey Brendan, you look gorgeous today. What beauty salon did you visit?’. is the concept ‘overwhelmingly tending towards a female rather than a male because of the words ‘gorgeous’ and ‘beauty salons’. Even though the lack of profanity or rudeness might give an impression of denoting positive affect, it relates more to females. If this comment was aimed at a boy, it might be an implicit way of accusing the boy of being effeminate, and thus a candidate sentence for cyberbullying.

An analysis of the comment shows that it is overwhelmingly more similar towards female concepts rather than male. The concepts of ‘gorgeous’ and ‘beauty salons’ are those that typically used in reference to girls rather than boys. If this comment was aimed at a boy, it might be an implicit way of accusing the boy of being effeminate, and thus becoming a candidate sentence for cyberbullying that deserves further scrutiny. Note that “gorgeous” by itself has a positive connotation, so it would be misinterpreted by something merely looking for positive vs. negative words.

Here, we have focused on LGBT accusations, but in much the same way, domain-specific knowledge about other topics connected with cyberbullying such as race and culture, intelligence and physical

appearance, social rejection etc can be built. Canonical concepts can be selected for each of the topics in much the same way. For example, for the topic of physical appearance, the concept of ‘fat’ would be a canonical concept. ‘*French fries*’ and ‘*cheeseburgers*’ for example, would be closer to the concept of ‘fat’ than ‘*salads*’.

We discuss the evaluation of both the statistical supervised learning methods and commonsense reasoning in section 4. An error analysis on the supervised learning methods highlights the need for commonsense reasoning. Of course, detection is just the first part of addressing cyberbullying. In the next part of this paper, we discuss some approaches for reflective user interaction and intervention to formulate an end-to-end model of tackling textual cyberbullying from detection to mitigation.

4. INTERVENTION STRATEGIES: REFLECTIVE USER INTERFACE

4.1 Monitoring and user privacy

Privacy advocates may object to having a detection algorithm scanning messages and text conversations, as this is a potential violation of the user’s privacy. Many common computing situations today involve the monitoring of user input. Users of the Google search engine and Gmail mail systems, for example, grant Google permission to analyze their mail or search terms in order to deliver targeted search results or targeted advertising. While many users are concerned about their privacy [48], others feel less concerned with having their input monitored by a program. In such cases it is the user’s responsibility to use the “opt-out” option, which may address their privacy concerns..

Many common computing situations today involve monitoring of user input. Users of the Google search engine and Gmail mail systems are granting Google permission to read their mail or search terms in order to deliver targeted search results or targeted advertising. While many users are concerned about their privacy [48], others feel less concerned with having their input monitored by a program. We recommend that users should always be given an “opt-out” option that can address their privacy concerns.

Minors, which have different privacy issues, are heavily engaged in the issues of cyberbullying [49]. Many parents insist on monitoring their children’s social interactions, and some establish behavioral rules for the use of social networks that are extremely restrictive. For younger children, some parents resort to social networks like “Scuttle Pad” [50] and “What’s What?” [51] which are promoted as “safe networks.” Similar websites prohibit any un-moderated commentary, any use of profanity, any social interaction with strangers, any reference to unapproved web sites, etc. New strategies in software-based intervention will hopefully contribute to an increased feeling of safety among parents and children, while still permitting considerable freedom of expression on the children’s part.

4.2 Roles in the bullying process

There are many roles in the cyberbullying process, which include the perpetrator, the victim and third party bystanders, such as friends, adults, moderators, and network providers. Each of these roles might elicit different kinds of reflective interfaces appropriate to their role. These roles are not mutually exclusive. Determining who is the victim and who is the perpetrator may not be an easy task. Victims may be tempted to cope with the situation by retaliating, which then thrusts them into the role of perpetrator [52]. In our collaboration with the social network provider Formspring [53], we learned that some negative interactions seem to start in one social network site, then spill onto another. Sometimes bullies in the digital realm may be victims in the physical world. Such complexity provides a source for misinterpretation of roles and behavior. Thus directly identifying an individual of being a perpetrator or a victim may not be constructive in diminishing negative behavior. True bullies may never be stymied by any intervention, real or digital, however providing tools to support healthy digital conversations is needed.

4.3 Reflective user interfaces

There are many non-computational challenges to the reliability of algorithmic detection of cyberbullying. People may legitimately differ in what they consider bullying. Seemingly humorous responses can become unknowingly hurtful. There may be cultural differences between users which cause miscommunication. And the context of a conversation may extend beyond the social network.

Given these challenges, careful consideration should be taken to plan the next actionable steps after being able to reasonably identify possible candidates for potentially bullying messages.

Reflective User Interface design is a novel approach to encouraging positive digital behavioral norms. Borrowing the framework from principles espoused by Donald Schön on Reflective Design [54], a Reflective User Interface is an array of solutions that might help stem, or change the spread of hurtful online behavior.

Schön stated three notions of the reflective practitioner, “reflection in action,” “reflection on action,” and “ladders of reflections.” One would reflect on behavior as it happens, so as to optimize the immediately following action. One reflects after the event, to review, analyze, and evaluate the situation, so as to gain insight for improved practice in the future. And one’s action, and reflection on action makes a ladder. Every action is followed by reflection and every reflection is followed by action in a recursive manner. In this ladder, the products of reflections also become the objects for further reflections. [55]

While most of their work refers to physical products, not software interfaces, researchers such as Sengers [56], Hallnäs, and Redström [57], also offer helpful insights in considering how to apply reflection to design.

Reflective User Interfaces include notifications, action delays, displaying hidden consequences, system suggested flagging, interactive education, and the visualization of aggregated data, addressing the challenges faced by both end-users and social network moderators. Through the interface, the end-user is encouraged (not forced) to think about the meaning of a given situation, and offered an opportunity to consider their options for reacting to it in a positive way. Reflection User Interfaces resist the urge to implement heavy-handed responses, such as direct censorship. Instead, the end-user is offered options to assist them to self-adjust or seek external help. For the network moderator, a dashboard interface tool to display high level network-wide overviews of aggregated negative user behavior, quickly identify problematic messages, and expedite actionable communications is in development.

4.4 End-user strategies

In providing tools to facilitate user discourse, users, moderators, and social network providers are encouraged to take an active part in determining and enforcing their own norms of healthy digital social behavior. These methods are not prescriptive, but rather options for social network providers to implement and test.

One class of interventions is notifications. Notification is drawing attention to a particular situation in order to encourage reflection. Oftentimes, people need only very subtle cues to help them understand how their behavior affects others. In face-to-face conversations in the real physical world, one has facial cues, body language etc. to help show how one’s input is being accepted. On phone calls, one can hear a person’s intonation, pitch, and volume. Based on these physical responses one can quickly adapt to curb or change their behavior. In the digital realm this is not the same, especially when conversations are not in real time. Changes in the user interface could make up some of these differences. [58]

Another class of interventions is interactive education. Current anti-cyberbullying efforts in schools and in youth-oriented media centers focus on educating participants about the process. [10] Most education efforts consist of general guidelines such as warning potential perpetrators about the negative consequences to their victims and the potential damage to their own reputations. They counsel potential victims to share their experiences with friends, family, and responsible adults. They counsel potential bystanders to recognize such situations in their social circle and to take steps to defuse the situation and to provide emotional support to the participants.

While these education efforts are positive contributions, they can be ineffective because they are disconnected from the particulars of the actual situation, both in relevance, and in time and space. Guidelines are often vague and they do not address the particular details of an actual situation.

Advice is usually so general that it is not directly actionable. The venue for bullying education is often school assemblies or classes, far from where bullying actually takes place.

The fact that cyberbullying occurs online gives an opportunity for intervention in real time. When a potential perpetrator is about to send a problematic message, there may be some time to encourage that person to reconsider, or to give them an opportunity to rescind their message. When a potential victim is about to receive a message, there may be a few minutes to counsel them on the appropriate response, or influence their immediate feelings about receiving such a message. Rather than give completely general advice, tailored advice may be offered addressing that particular kind of bullying. Such advice can be immediately actionable, and can have a dramatic effect on the outcome of the situation.

4.5 Introducing action delays

A number of possible intervention techniques are aimed, not at interrupting the process, but at introducing small delays to the process in the hopes that the delay will encourage reflection. Such delays may not prevent severe cases of bullying, however major cyberbullying problems are often characterized by rapid spread in a particular community. Slowing down the spread of negativity might in some circumstances be enough to avert a major disaster. [59] The aim is to slow the spread below the “chain reaction” rate.

Alerting the end-user that their input might be hurtful and making them wait some time before actually submitting could also be helpful. The end-user could decide to rephrase their comment or cancel it outright. This enforces a time for Schön’s “reflection in action.” Generally user interface design has been focused on helping the end-user get or submit information as quickly as possible. However there are cases where offering the user time to reconsider and confirm that the information they are providing is truthful is warranted, as in credit card purchases, applications, etc. Such enforced reflection is also common on commercial sites, which provide free services to a user. Web sites such as RapidShare, the file-sharing site [60] enforce a delay so that the user has time to consider the worth of the service, and the value of purchasing enhanced services or delay-free usage.

Even after making the decision to send the message, it is also helpful to provide a delay before the message is actually delivered to the recipient, giving the user the opportunity to undo the action and take the message back before it is seen by the recipient. Often the act of sending makes the consequences of sending seem more real to the user, and triggers a “sender’s remorse” response.



Figure 6 – Mock-up of delay and undo operations given to the sender for a chance to reconsider message.

4.6 Informing the user of hidden consequences

Oftentimes the end-user does not realize that they are responding to the group’s entire social graph, not just to the owner of the page they are commenting on. Whether or not a single comment, or the overall tone of thread, is deemed negative by the detection algorithm, an interface change to the text label on the submit button may reflect the number of people they are communicating with.

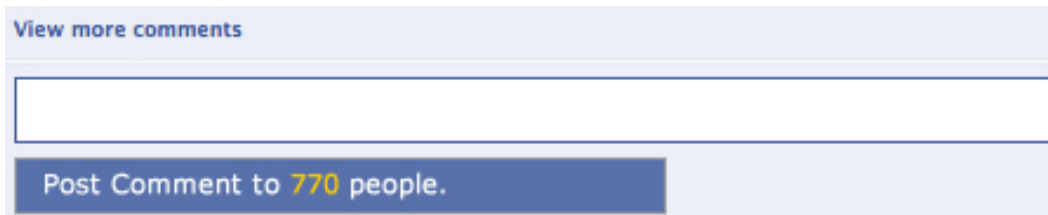


Figure 7 Mock-up of informing the sender of the consequences of sending to a large social network.

For example, if Romeo, who has 350 friends, is posting a comment on Juliet's page, who has 420 friends, then the submit button for Romeo would reflect, "Send to 770 people."

In another example, if Tybalt's comment on Juliet's page is negative, after successfully submitting, the system might respond with an alert box, "That's sounds harsh! Are you sure you want to send that?" If Tybalt changes his mind, an "undo" button could be made available, as his comment has yet to be sent.

4.7 Suggesting educational material

For Juliet, the receiver of negative comments, the interface could provide interactive educational support. Using Google Gmail ad-like text messages next to the negative comments offering the user support. "Wow! That sounds nasty! Click here for help." The small text message links to external websites related to supporting victims of bullying. This also provides an easy conduit for external support agencies to connect their materials to the individuals for whom they work. Social network providers could partner with outside social agencies to craft appropriate material to link to, and utilize this method of helping their end-users.



Figure 8 – Mock-up of a small text message offering educational material to users, after detection of a problematic message.

We agree that there is no guarantee that bullies will opt-in to clicking on educational material, nor that such educational material would have a significant effect on stopping bullying behavior. But there is evidence that some kinds of educational material are indeed effective against bullying. The opt-in would at least give such material an opportunity to reach its intended target in a situation close to the time and place where effective action is possible. The results of our user study provide some evidence that enough participants found the links useful that make it plausible that the links would be clicked on, though this would remain to be verified in field tests.

We do cite evidence on the effectiveness of anti-bullying education programs [73] that include educational material in the form of videos, school assembly presentations, and readings on the subject. It states, "Our meta-analysis showed that, overall, school-based anti-bullying programmes are effective in reducing bullying and victimization. The results indicated that bullying and victimization were reduced by about 17–23% in experimental schools compared with control schools..." "The most important programme elements that were associated with a decrease in victimization (i.e. being bullied) were videos..."

Educational materials created to support victims of bullying are often too general. And the actual support provided to victims usually happens long after the bullying event. Even more ambitious than a link to external content, we are building an interface strategy to provide short targeted video suggestions of what to do next, or how to respond. The detection algorithm could analyze the user's comments and situation, and then align it to preselected stories or educational materials representing similar issues.

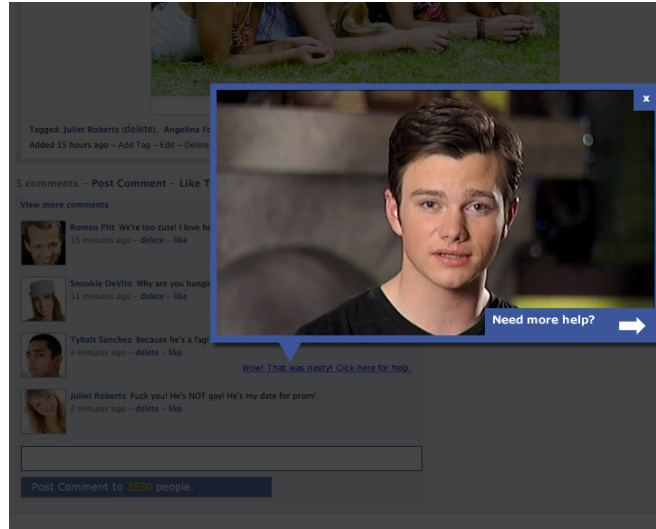


Figure 9 Mock-up of an anti-bullying video from <http://www.itgestbetter.org> is presented, after the user clicks on a help message link.

The small text link is the first point of entry for the user. The interface is developed to include functionality similar to an expert help system [61]. Designing an intelligent help system to best serve the end-user is complex and difficult, as discussed by Molich and Nielsen. [62] The text link initiates the end-user “reflection on action.” After the educational material is presented, the interface could ask the user whether or not the story provided was in fact a good match and if it was useful. If the user requests more help, suggested solutions and materials are provided in a tiered method. The help support system would record the user interactions, so that if the user requests more help in the future, the system knows it has provided assistance before, and would not treat the interaction as a new occurrence. By allowing the user to opt in or out at any stage of engagement, the support would become contextual, prescriptive, and desired, rather than overbearing and obstructive.

4.8 Social network provider strategies

The primary concern of social network providers and group moderators is to provide a safe and welcoming environment for their community. It is not necessary to detect or control every single incident of bullying that occurs. Most important is to prevent an initially trivial incident from escalating, or spreading within a community from becoming the social norm. An important part of maintaining quality social networks would include giving moderators an aggregate view of the behavior of their end-users. Patterns of escalation or spreading of bullying caught early give moderators opportunities to intervene before serious problems arise.

4.9 Flagging of messages

Many social networks allow end-users to flag messages as inappropriate. Human moderators, who use their judgment to decide whether action is warranted, often review such flagged messages. Automated detection algorithms can contribute to this flagging process by suggesting to a participant that a message might need to be flagged. It can also help by prioritizing a list of flagged messages for review by the human moderator. This prioritization is essential, because the volume of flagged messages in large-scale social networked can overwhelm moderators.

Flagged comments may be displayed in various ways. The comment may be visibly marked or hidden to the public, particular end-users, or available for viewing to only the receiver and sender of the comment. The moderator could also hold a flagged comment for review.

4.10 Visualization of Online Bullying

A dashboard interface for moderators and social network providers is in development to visualize high-level overviews of their network’s end-user behavior. One view of the dashboard could serve some of the same functions as back channels [63] in calling attention to possibly problematic situations. A

social network using a detection algorithm may not want to make any changes to the end-user interface until after understanding the scope and domains of their end-users' negative behavior.

The dashboard would have many display views to reflect key semantic terms, social clusters, events, and basic demographics derived from the network social graph. It could also display the actual flagged messages related to the semantic terms, prioritized in order of seriousness by the detection algorithm. Providers could use the dashboard to help their moderators get an overview of their supervised space on the network.

Problems in the real world are often reflected in the digital. [64] As a courtesy service, the social network provider could also provide third parties, such as police and school administrators/staff a version of the dashboard using sanitized (anonymous) user names. In this scenario, a school administrator would be able to see an overview of the digital behavior of their school's student population. For example, without giving actual screen names, and real conversations, the school could find out that there are problems such as gay bashing during the weeks leading up to the prom. This information could be vital in scheduling appropriate real world intervention strategies at the school.

Due to the public's growing hypersensitivity of cyberbullying, sometimes one publicized incident can over-represent the severity of problems in a social network. [65] A public version of the dashboard could provide transparency and a more balanced overview about the network's problems, based on objective data.



Figure 10 – Mock-up of social network dashboard displaying the community visualization environment.

5. EVALUATION AND DISCUSSION

The statistical models discussed in section 3 were evaluated against 200 unseen instances for each classifier. The labels assigned by the models were compared against the labels that were assigned to the instances during annotation. The accuracy and kappa values of the classifiers are in the tables below.

To avoid lexical overlap, the 200 instances for each label were derived from video comments that were not part of the original training and validation data. Prior work on the assessment of classifiers suggests that accuracy alone is an insufficient metric to gauge reliability. The kappa statistic κ (Cohen's kappa), which takes into account agreement by chance, has been argued as a more reliable metric in conjunction with accuracy [44]. We evaluate each classifier in terms of the accuracy, the F1-score, as well the kappa statistic.

Multiclass classifiers underperformed compared to binary classifiers. In terms of accuracy, JRip was the best, although the kappa values were best with SVM. SVM's high kappa values suggest better reliability for all labels. Naïve Bayes classifiers for all labels perform much better than J48.

	Naïve Bayes			Rule-based Jrip			Tree-based J48			SMO (SVM)		
	Accuracy	F1	κ	Accuracy	F1	κ	Accuracy	F1	κ	Accuracy	F1	κ
Sexuality	66%	0.67	0.657	80.20%	0.76	0.598	63.40%	0.57	0.573	66.70%	0.77	0.79
Race	66%	0.52	0.789	68.30%	0.55	0.789	63.50%	0.48	0.657	66.70%	0.63	0.718
Intelligence	72%	0.46	0.467	70.39%	0.51	0.512	70%	0.51	0.568	72%	0.58	0.772

Figure 11 - Binary classifiers for individual labels

Mixture	63%	0.57	0.445	63%	0.60	0.507	61%	0.58	0.456	66.70%	0.63	0.653
---------	-----	------	-------	-----	------	-------	-----	------	-------	--------	------	-------

Figure 12 – Multiclass classifiers for the merged dataset – Binary classifiers trained for individual labels fare better than multiclass classifiers trained for all the labels. JRip gives the best performance in terms of accuracy, whereas SMO is the most reliable as measured by the kappa statistic.

5.1 Error analysis of the supervised learning models

As we hypothesized, an error analysis on the results reveals that instances of bullying that are apparent and blatant are simple to model because of their stable, repetitive patterns. Such instances either contains commonly used forms of abuse or profanity, or expressions denoting a negative tone. For example, consider the following instances:

u1 as long *as fags don't bother me* let them do what they want

u2 hey we didn't *kill* all of them, some are still alive today. And at least we didn't *enslave them* like we did the *monkeys*, because that would have been more humiliating

Both the instances shown above (the first pertaining to sexuality and the second pertaining to race) contain unigrams and expressions that lend them to be positively classified by the models. Instances such as the ones shown above, which contain lexical and syntactic patterns of abuse, lend themselves to supervised learning for effective detection. However, the learning models misclassified instances that do not contain these patterns and those that require at least some semantic reasoning. For example, consider the following instances:

u3 they make beautiful girls, especially the one in the green top

u4 she will be good at pressing my shirt

In the first instance, which was posted on a video of a middle school skit by a group of boys, the bully is trying to ascribe female characteristics to a male individual in the video. The instance has no

negativity or profanity, but implicitly tries to insult the victim by speculating about his sexual orientation. ‘Tops’ and ‘beautiful’ are concepts that are more associated with girls rather than boys, and hence if attributed to the wrong gender, can be very hurtful. In the second instance, a bully exploits the common sexist stereotype that pressing clothes is an act reserved primarily for women. The learning models misclassified these two instances, as it would need to have some background knowledge about the stereotypes and social constructs and reason with it. In the next section, we discuss our work with supervised learning models in the context of related approaches to sentiment analysis.

5.1.1 Discussion

Prior research in sentiment analysis has focused on sentiment polarity for opinion analysis for movie and product reviews [45]. However, the nature of interpersonal and group interaction on social networks is different from sentiment polarity of reviews from two perspectives and hence difficult to compare them. First, interaction of social networks (like Formspring) as a sociolinguistic phenomenon is more targeted towards a specific audience (an individual or a group of individuals), whilst movie and product reviews are intended for a larger, more general audience. Second an analysis of discourse on social networks involves deeper attributes such as identity, ascription to a particular community, personality and affect [46], which is more than just sentiment polarity of movies or product reviews where there is a prior acknowledgement of the domain under scrutiny (a movie or a particular product).

Recent work with affect recognition in text has attempted a fine-grain compositional approach to gauging emotions [47]. While we did not adopt a finer granularity approach towards gauging emotion, we emphasize that our focus was on overcoming the limitations of supervised learning methods in catching indirect, subtle forms of abuses using social constructs which requires reasoning along relevant dimensions (such as gender roles). In the next section, we discuss the evaluation of the experiments involving commonsense reasoning.

5.2 Evaluation of commonsense reasoning models

For an evaluation of the detection part involving commonsense reasoning, it is essential to have a dataset that contains instances of cyberbullying that are devoid of profanity and are implicitly crafted to insult or malign a user. To address the specific slice of LGBT bullying in this work, it is essential to have a test dataset that pertains to LGBT bullying as well as some instances that do not pertain to LGBT issues.

We build such a test by performing a filtering operation on the original Formspring dataset as follows: the same set of people who annotated the YouTube corpus were asked to pick instances from the Formspring dataset that satisfied the dual criteria of not having any profanity and implicitly trying to attack, insult or speculate the sexuality of the victim. Of the 61 instances of bullying that were obtained from the three annotators, 50 instances were made into a test set. It is important to keep in mind that the original Formspring corpus contains instances that have already been flagged as bullying. Hence the annotators were not asked to check if an instance was bullying or not.

Since the goal of this detection approach that we take in this paper is to prioritize reported instances of bullying based on similarity scores, we adopt a similar approach for this test dataset. The test dataset was evaluated with the approach mentioned in section 2 to generate similarity metrics for each instance with the canonical concepts *girl* and *boy*. The results were shown to each of the three annotators to check if they agreed with the metrics generated by the commonsense reasoning model. The results are as follows:

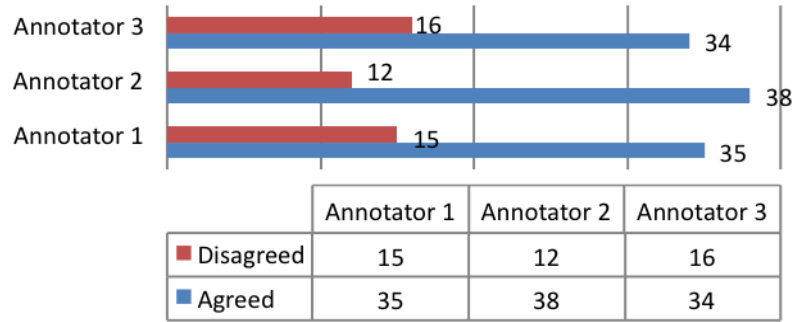


Figure 13 – Evaluation of the commonsense reasoning model. 50 instances from the Formspring dataset were evaluated by the model to generate similarity scores for the canonical concepts girl and boy. The same three annotators, as shown above, validated the results. All the 50 instances were previously flagged as instances of cyberbullying.

5.2.1 Error Analysis of the commonsense reasoning model

An analysis of the instances for which the annotators disagreed with the commonsense reasoning model can be classified into three kinds. The first kind were instances for which the similarity metrics did not make common sense and the second kind were instances for which the annotators did not agree to the scale of the similarity metrics.

Most of the instances for which similarity metrics did not make commonsense were largely due to sparsity of data in the space that was built for performing the SVD. For instance, consider the following example having similarity metrics that did not make common sense:

George Michael or Elton John?

This instance received an extremely high score for the concept “boy” owing to the names of the individuals mentioned above. However, a deeper analysis shows that the above individuals are celebrity singers who also have one thing in common between them – they are both openly gay. The three annotators all agreed that by suggesting that an individual likes these singers, the perpetrator is implicitly trying to speculate or mock their sexuality. To address such instances, one really needs to have more canonical concepts than *girl* and *boy*.

Those instances for which the relative scale of the similarity scores was not agreeable to the annotators can be attributed to the scoring process in ConceptNet, which relies on the frequency of an assertion to determine its relative scoring []. For example, consider the following instance:

why did you stop wearing makeup?

The above example generated a normalized similarity scores as follows: 60.7% for the concept of girl and 32.2% for the concept of boy. While there can be men who in various roles such as actors routinely wear makeup, makeup is more strongly associated with women than men. This suggests the need for an in-context weighting of assertions. Makeup and costumes, for example are more likely to be associated with individuals in the performing arts, irrespective of their gender.

It is clear from the evaluation that the problem of sparsity, as well as the ability to individually weight an assertion will be vital if this approach is to be implemented in a large-scale user community. One can imagine crowd-sourced collection of such relevant social constructs as commonsense assertions from both users and moderators of social networks.

5.3 Reflective User Interface Evaluation

The Reflective User Interface strategy of suggesting educational material, as discussed in section 4.7, was evaluated in a small user study, testing the differences between dynamic in-context targeted

advice in the user interface, targeted static advice in the user interface, and the typical “help” link user interaction found on most social networks. The study included five participants, consistent with the findings Nielsen suggests when conducting a user test. [66] A fully functioning hypothetical social network, called Fakebook, was built as platform for testing both detection algorithms and user interface strategies. Fakebook was graphically styled and patterned after the popular social network Facebook, so that users would feel familiar with its interface.



Figure 14 – Fakebook, the fully functioning social network built as a testing platform. The “Wall” interface is shown with in-context links to targeted help.

Prior to the user study, a mock conversation between three imaginary persons was staged to present a bullying interaction. While the dialogue was fictional, it modeled real messages found from research data provided by MTV’s A Thin Line website. Using the detection classifier, each message was scanned for being a possible case of bullying.

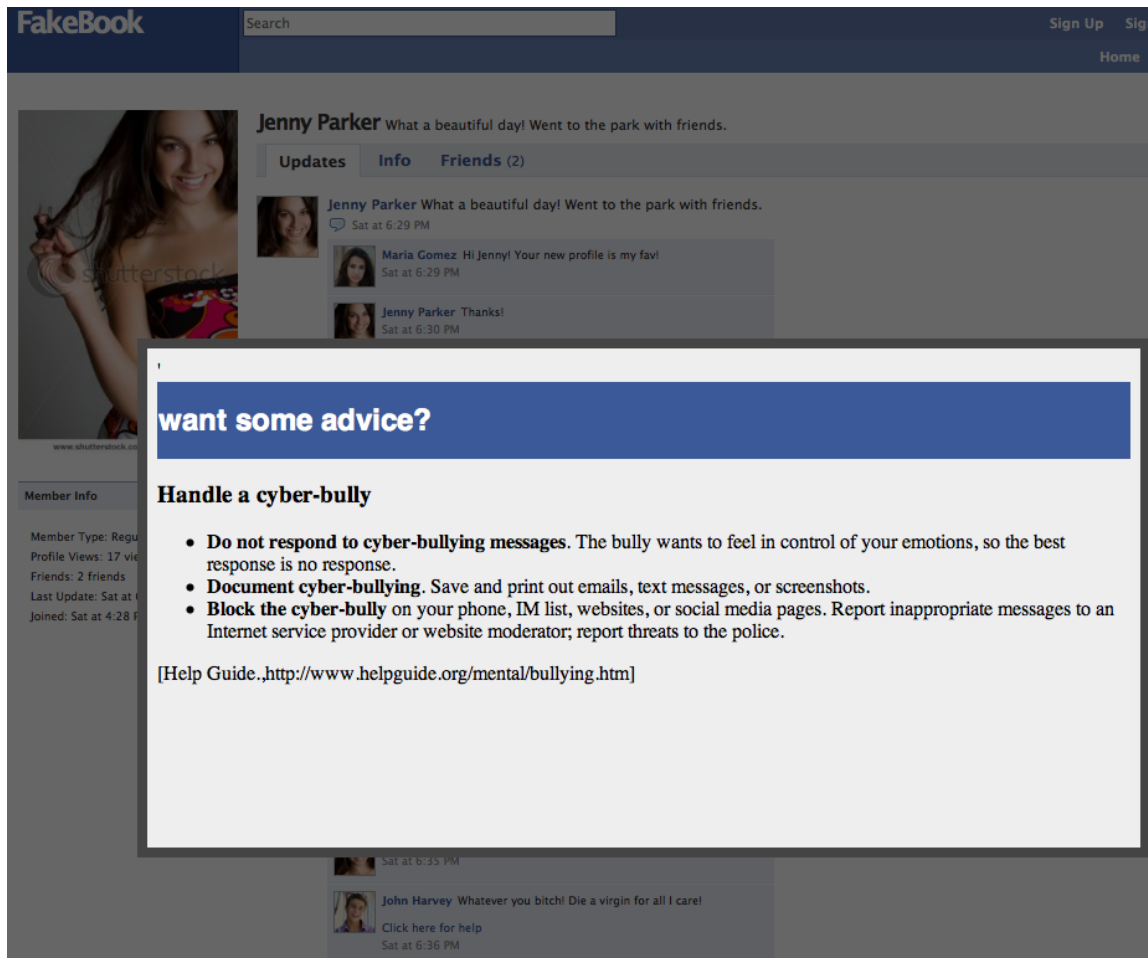


Figure 15 – After the end-user clicks on a link for more help, the Fakebook modal window displays in-context targeted help.

Three different versions of a “wall” (an interface that allows users to send and receive messages) [67] conversation were created, each varying the type of advice offered by the user interface. In the first version, a small text link stating, “Click here for help,” was placed next to the messages positively identified by the classifier as being a candidate for bullying. Once clicked, the link would bring up a modal window (pop-up window) containing a short paragraph of advice for coping with bullying situations. While the advice was hand-curated from a website [68] specializing in cyberbullying coping advice for both children and parents, it is dynamic displayed based on the detection algorithm’s analysis of the bullying message. Jakob Nielsen’s Ten Usability Heuristics provides a relevant guide for the modal window interface design decision. [69] “Recognition rather than recall” suggests that, “the user should not have to remember information from one part of the dialogue to another.” The interface presents the help advice in the same viewing area as the potentially negative interaction.

The second version of the interface looked exactly as the first, but the content of the “Click here for help” modal window consisted of a single web link to a website [70]. The suggested website represents many similar websites that are listed in the help sections of many social networks. These links, while helpful, are often hard for end-users to find. And they are located on webpages separate from the user interaction space. The third version of the interface contained no targeted advice links. In every version, the standard “help” link was present. Clicking this link brought users to a page mirroring the current Facebook help page. [71]

- Basics
- Trouble Using Facebook
- Report Abuse or Policy Violations**
- Ads and Business Solutions
- Safety Center
- Community Forum

Report Abuse or Policy Violations

English (US) ▾

Report Abuse or Policy Violations

Expand All · Share

Tools for Addressing Abuse

- Tools for addressing abuse
- How to report abuse
- Privacy settings
- Tools for reporters without accounts

Report Abuse or Policy Violations

- Impostor accounts
- Bullying
- Intellectual property infringements
- Unauthorized payments
- Advertising violations
- Pornography
- Scams, phishing and spam
- Violent, graphic or gory content
- Hate speech
- Promotion of "cutting," eating disorders or drug use

Secure a Compromised Account

- Hacked, scammed or phished accounts
- Best practices for account security

Surface Safety and Privacy Concerns

- Memorializing accounts of deceased friends or family
- Reporting underage accounts
- Helping someone who posts suicidal content
- Reporting convicted sex offenders

Get Help and Give Feedback

- Family Safety Center
- Post and answer questions in the Help Forum
- Report a bug if something isn't working

Figure 16 – Facebook’s help page.

For the user study, participants used the Fakebook social network, to take a survey. They were shown the three versions of the wall of Jenny, a fictional character, and her conversation with two persons, John and Maria. Participants were asked to read through the conversational thread imagining themselves as each of one of the characters, and then asked questions. The five participants were asked to click on each “*Click here for help*” link while reading through the conversation. Participants were told that Jenny was the victim, John was the bully, and Maria was a third-party bystander.

The survey used Likert scale questions, answering with Strongly Disagree, Disagree, Neither Agree nor Disagree, Agree, or Strongly Agree. Participants were asked the same three questions for each version of the wall interface. The first question was, “Imagine you are Jenny. Assuming Jenny is the victim, when I clicked on the advice links I considered the advice helpful in the situation.” The second question, “Imagine you are John. Assuming John is the bully, when I clicked on the help links, I felt reflective about my behavior and how it might have affected Jenny.” And the third question, “Imagine you are Maria. Assuming the Maria is a bystander, when I clicked on the links, I reflected on how the messages might have affected Jenny.”

	Interface 1: In-Context Dynamic Help	Interface 2: (Control) Static Help	Interface 3: No Interface Change
Motivations	Targeted help is more appropriate to the situation of the end-user. By providing the help in the same	Static help provided to the end-user in the same interface as the bullying interaction, would have more perceived value to	This represents the standard social network interface.

	interface as the bullying interaction, the advice becomes actionable.	the end-user than no in-interface assistance at all.	
User Protocol	Participants were asked to click help links.	Participants were asked to click help links.	Participants were asked to click the standard 'help' link.

Figure 17 – Table describing study protocol.

Interface 1: In-Context Dynamic Targeted Help					
	Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
Imagine you are Jenny. Assuming Jenny is the victim, when I clicked on the advice links I considered the advice helpful in the situation.	0%	0%	0%	20%	80%
Imagine you are John. Assuming John is the bully, when I clicked on the help links, I felt reflective about my behavior and how it might have affected Jenny.	0%	20%	0%	40%	40%
Imagine you are Maria. Assuming the Maria is a bystander, when I clicked on the links, I reflected on how the messages might have affected Jenny.	0%	0%	0%	20%	80%

Figure 18 – Table showing results for interface 1 using in-context, targeted dynamic help.

Interface 2: (Control) Static Help					
	Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
Imagine you are Jenny. Assuming Jenny is the victim, when I clicked on the advice links I considered the advice helpful in the situation.	0%	40%	40%	20%	0%
Imagine you are John. Assuming John is the bully, when I clicked on the help links, I felt reflective about my behavior and how it might have affected Jenny.	20%	60%	20%	0%	0%

Imagine you are Maria. Assuming the Maria is a bystander, when I clicked on the links, I reflected on how the messages might have affected Jenny.	0%	80%	20%	0%	0%
---	----	-----	-----	----	----

Figure 19 – Table showing results for interface 2 using static help.

Interface 3: No Interface Changes					
	Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
Imagine you are Jenny. Assuming Jenny is the victim, when I clicked on the advice links I considered the advice helpful in the situation.	100%	0%	0%	0%	0%
Imagine you are John. Assuming John is the bully, when I clicked on the help links, I felt reflective about my behavior and how it might have affected Jenny.	100%	0%	0%	0%	0%
Imagine you are Maria. Assuming the Maria is a bystander, when I clicked on the links, I reflected on how the messages might have affected Jenny.	100%	0%	0%	0%	0%

Figure 20 – Table showing results for interface 3, which contained no interface changes.

We were encouraged that participants overwhelmingly preferred the interface with targeted in-context advice, which concurs with our assertion that targeted help within the user interface at the point of the bullying interaction would be more helpful to the end-user, than the typical “help” link support provided by social networks. There are few, if any, social networks providing in-context dynamic support for cyberbullying. As a result there are few intervention models to employ in the manner we propose. The second interface providing static help within the user interaction could serve as intermediate step towards providing end-users with support.

6. CONCLUSION

Cyberbullying of youth on social networks is a growing problem, as recent news stories detailing suicides of bullied teenagers attest. In less extreme forms, this problem is widespread, with a significant fraction of youth reporting bullying incidents happening to them or to their social circle. Such incidents threaten the continued growth of online social networks for young people. This would be a shame, as social networks provide many benefits for youth, including opportunities for forming new relationships, strengthening existing ones, sharing interests, and practicing reading and writing in a personally meaningful context.

In addition to educational efforts in schools, increasing awareness, and discussion of the problem between adults and young people, we believe technical solutions in social network software will form an important component of identifying and reducing the harm associated with this problem. We have

presented a suite of capabilities for social network software that address detection of potentially bullying messages, and intervention by notifying participants and network moderators, managing message access, and offering targeted educational material.

Fully general natural language understanding still remains beyond reach. But we have shown that state-of-the-art natural language processing, augmented by a Commonsense reasoning technique, specialized knowledge bases concerning bullying, and a novel reasoning technique can result in accuracies approaching 80% on identifying potentially bullying messages with significant agreement ratings with human labelers.

We have also presented a wide range of potential intervention techniques, ranging from subtle changes to the messaging interface, to personalized, interactive educational material, and “air traffic control” overviews to help network providers maintain positive social norms. These interfaces show how adding intelligence to an interactive interface, in cooperation with the various roles that human users play, can make social network applications more effective in their goal of enhancing human relationships.

ssss

REFERENCES

1. Ybarra, M. Trends in technology-based sexual and non-sexual aggression over time and linkages to non-technology aggression. Presentation at the National Summit on Interpersonal Violence and Abuse Across the Lifespan, 2010: Forging a Shared Agenda. Houston, TX.
2. Vandebosch H., Cleemput, K.V Cyberbullying among youngsters: profiles of bullies and victims *New Media & Society* December 2009 11: 1349-1371, first published on November 24, 2009.
3. Menesini, E., & Nocentini, A. (2009). Cyberbullying definition and measurement: Some critical considerations. *Journal of Psychology*, 217, 230-232.
4. Sourander, A., Klomek, A.B., Ikonen, M., Lindroos, J., Luntamo, T., Koskelainen, M., Ristkari, T., Helenius, H. Psychosocial Risk Factors Associated With Cyberbullying Among Adolescents: A Population-Based Study *Arch Gen Psychiatry*, Jul 2010; 67: 720 – 728.
5. Cyberbullying, The National Crime Prevention. Retrieved from <http://www.ncpc.org/cyberbullying>.
6. boyd, danah. “Why Youth (Heart) Social Network Sites: The Role of Networked Publics in Teenage Social Life.” *MacArthur Foundation Series on Digital Learning – Youth, Identity, and DigitalMedia* Volume (ed. David Buckingham). Cambridge, MA: MIT Press.
7. The White House Conference on Bullying Prevention, Press Release. Retrieved from <http://www.whitehouse.gov/the-press-office/2011/03/10/background-white-house-conference-bullying-prevention>.
8. The White House led interdepartmental initiative on proving resources to stopping bullying. Retrieved from http://www.stopbullying.gov/references/white_house_conference/.
9. Kontostathis, A., Edwards, L. and Leatherman, A. (2010) Text Mining and Cybercrime, in *Text Mining: Applications and Theory* (eds M. W. Berry and J. Kogan), John Wiley & Sons, Ltd, Chichester, UK.
10. The Jay Banks anti-bullying program. Retrieved from <http://jaybanks.com/anti-bullying-program/>.
11. A Thin Line, MTV's sexting, cyberbullying, digital dating abuse campaign. Retrieved from www.athinline.org
12. Mishna, F., Saini, M., Solomon, S., Ongoing and online: Children and youth's perceptions of cyber bullying, *Children and Youth Services Review*, Volume 31, Issue 12, December 2009, Pages 1222-1228.

13. Patchin, J. W. & Hinduja, S. (2012). *Cyberbullying Prevention and Response: Expert Perspectives*. New York: Routledge.
14. Greevy, E., Smeaton, A.F., Classifying racist texts using a support vector machine. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '04)*. ACM, New York, NY, USA, 468-469.
15. Dinakar K., Reichart R., Lieberman, H., Modeling the detection of textual cyberbullying, *International Conference on Weblog and Social Media - Social Mobile Web Workshop*, Barcelona, Spain 2011
16. Ott, M., Choi, Y., Cardie, C., Hancock, J.T. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1 (HLT '11)*, Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 309-319.
17. Chin, S., Street, N., Srinivasan, P., Eichmann, D., 2010. Detecting Wikipedia vandalism with active learning and statistical language models. In *Proceedings of the 4th workshop on Information credibility (WICOW '10)*. ACM, New York, NY, USA, 3-10.
18. Vala, M., Sequeira, P., Paiva, A., Aylett, R. FearNot! demo: a virtual environment with synthetic characters to help bullying. In *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems (AAMAS '07)*.
19. Olweus D. *Bullying at School: What We Know and What We Can Do*. Oxford, England: Blackwell; 1993.
20. Mishna F, Cook C, Gadalla T, Daciuk J, Solomon S. Cyber bullying behaviors among middle and high school students. *Am J Orthopsychiatry*. 2010;80(3):362-374.
21. Blumemenfeld, W.J, Cooper, R.M. LGBT and Allied Youth Responses to Cyberbullying: Policy Implication. *International Journal of Critical pedagogy* 2010; 3 (1): 114-133.
22. Heck, N.C., Flentje, A., Cochran, B.N., Offsetting risks: High school gay-straight alliances and lesbian, gay, bisexual, and transgender (LGBT) youth. *School Psychology Quarterly*, Vol 26(2), Jun 2011, 161-174.
23. Hinduja, S., & Patchin, J. W. (2008). Cyberbullying: An exploratory analysis of factors related to offending and victimization. *Deviant Behavior*, 29, 129–156.
24. Bramsen, P., Escobar-Molano, M., Patel, A., Alonso, M., 2011. Extracting social power relationships from natural language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1 (HLT '11)*, Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 773-782.
25. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, H. The WEKA data mining software: an update. *SIGKDD Explor. Newsl.* 11, 1 (November 2009), 10-18.
26. Cohen, W.W., Singer, Y. A simple, fast, and effective rule learner. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, 1999.
27. Sasaki, M.; Kita, K. Rule-based text categorization using hierarchical categories, 1998 IEEE International Conference on Systems, Man, and Cybernetics, 1998. vol.3, no., pp.2827-2830 vol.3, 11-14 Oct 1998.
28. Quinlan, R. *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, CA, 1999.
29. T. Gonçalves and P. Quaresma. A preliminary approach to the multilabel classification problem of Portuguese juridical documents. In F. Moura-Pires and S. Abreu, editors, *Proceedings of the 11th Portuguese Conference on Artificial Intelligence - EPIA '2003*, LNAI 2902, pages 435--444, Évora, Portugal, December 2003.
30. Cortes, C., Vapnik, V. "Support-Vector Networks", *Machine Learning*, 20, 1995, available at <http://www.springerlink.com/content/k238jx04hm87j80g/>
31. Tong, S., Koller, D., Support Vector Machine Active Learning with Applications to Text Classification, *Proceedings of the Seventeenth International Conference on Machine Learning*, p.999-1006, June 29-July 02, 2000.
32. Rogati, M., Yang, Y. High-performing feature selection for text classification. In *Proceedings of the eleventh international conference on Information and knowledge management (CIKM '02)*. ACM, New York, NY, USA, 659-661.

33. Gabrilovich, E., Markovitch, S. 2004. Text categorization with many redundant features: using aggressive feature selection to make SVMs competitive with C4.5. In Proceedings of the twenty-first international conference on Machine learning (ICML '04). ACM, New York, NY, USA, 41-.
34. Joachims, T. 1998. Text categorization with support vector machines: learning with many relevant features. In Proceedings of ECML-98, 10th European Conference on Machine Learning (Chemnitz, DE, 1998), pp. 137–142.
35. Gao, Y., Sun, S. "An empirical evaluation of linear and nonlinear kernels for text classification using Support Vector Machines," Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference on , vol.4, no., pp.1502-1505, 10-12 Aug. 2010.
36. Ortony, A., Clore, G.L., Foss, M.A, The Referential Structure of the Affective Lexicon. *Cognitive Science* 11(3): 341-364 (1987).
37. Singh, P., Lin, T., Mueller, T., Lim, G., Perkins, T., Zhu, W.L., Open Mind Common Sense: Knowledge Acquisition from the General Public, On the Move to Meaningful Internet Systems, 2002 - DOA/CoopIS/ODBASE 2002 Confederated International Conferences DOA, CoopIS and ODBASE 2002, p.1223-1237, October 30-November 01, 2002.
38. Mey, J. 2001. *Pragmatics: An Introduction*, page 76-77. Blackwell.
39. Liu, H., Singh, P., 2004. ConceptNet - A Practical Commonsense Reasoning Tool-Kit. *BT Technology Journal* 22, 4 (October 2004), 211-226.
40. Speer, R., Havasi, C., Lieberman, H. 2008. AnalogySpace: reducing the dimensionality of common sense knowledge. In Proceedings of the 23rd national conference on Artificial intelligence - Volume 1 (AAAI'08), Anthony Cohn (Ed.), Vol. 1. AAAI Press 548-553.
41. Havasi, C., Speer, R., Pustejovsky, J., Lieberman, L. Digital Intuition: Applying Common Sense Using Dimensionality Reduction. *IEEE Intelligent Systems*, October 2009.
42. Cambria, E., Hussain, A., Havasi, C., Eckl, C. AffectiveSpace: Blending Common Sense and Affective Knowledge to Perform Emotive Reasoning. *Workshop on Opinion Mining and Sentiment Analysis*, November 2009.
43. Cambria, E., Speer, R., Havasi, C. SenticNet: A Publicly Available Semantic Resource for Opinion Mining. *Proceedings of the AAAI 2010 Symposium on Common Sense Knowledge*.
44. Carletta, Jean. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249-254.
45. Pang, B., Lee, L. *Opinion Mining and Sentiment Analysis, Foundations and Trends in Information Retrieval*, v.2 n.1-2, p.1-135, January 2008.
46. Ginger, J., *Gender Roles & Group Discourse – The Facebook Project*, Retrieved from http://www.thefacebookproject.com/research/jeff/publications/digitaletnography/genderrolessroupdiscourse_rev1.pdf.
47. Neviarouskaya, A., Prendinger, H., and Ishizuka, M. Compositionality principle in recognition of fine-grained emotions from text. In *Proc. ICWSM 2009*, AAAI Press (2009), 278--281.
48. Electronic Privacy Information Center. (2011). *Google Privacy FAQ*. Retrieved from <http://epic.org/privacy/gmail/faq.html>.
49. *Teens and Cyberbullying. Executive Summary of a Report on Research*. Retrieved from <http://www.npc.org/resources/files/pdf/bullying/Teens%20and%20Cyberbullying%20Research%20Study.pdf>.
50. Scuttle Pad. (2011). Retrieved from at <http://www.scuttlepad.com/>.
51. *What's What?* (2011). Retrieved from from <http://whatswhat.me/>.
52. *Stop Cyberbullying*. (2011). *How to handle a cyberbully*. Retrieved from <http://www.stopcyberbullying.org/parents/howdoyouhandleacyberbully.html>.
53. *Formspring*. (2011). Retrieved from <http://formspring.me>.
54. Schön D., *The Reflective Practitioner*. Basic Books: New York, 1983.
55. *Learning & Computing Education: Reflections and Ideation*. (2011). *Design is a Reflective Practice: A Summary of Schön's Views*. Retrieved from <http://goelsan.wordpress.com/2010/08/20/design-is-a-reflective-practice-a-summary-of-schons-views/>.

56. Sengers, P. et al.(2005) Reflective Design. In Proceedings of the 4th decennial conference on Critical computing, pp. 49-58. ACM Press.
57. Hallnäs,L. & Redström,J. (2001) Slow Technology; Designing for Reflection. In: Personal and Ubiquitous Computing, Vol.5, No.3, pp.201 212. Springer.
58. Walther, J.B., Loh, T., & Granka, L. Let me count the ways: The interchange of verbal and nonverbal cues in computer-mediated and face-to-face affinity. *Journal of Language and Social Psychology*, 24, 36--65, 2005.
59. Madlock, P., & Westerman, D. (2011). Hurtful cyber-teasing and violence: Who's laughing out loud?. *J Interpers Violence*, 26(17), 3542-3560.
60. Rapidshare. (2011). Retrieved from <http://rapidshare.com>.
61. Ignizio, J., (1991). *Introduction to Expert Systems*. McGraw-Hill Companies.
62. Molich, R., Nielsen, J. 1990. Improving a human-computer dialogue. *Commun. ACM* 33, 3 (March 1990), 338-348.
63. McNely, B. 2009. Backchannel persistence and collaborative meaning-making. In Proceedings of the 27th ACM international conference on Design of communication (SIGDOC '09). ACM, New York, NY, USA, 297-304.
64. Hinduja, S., Patchin, J. 2007. "Offline Consequences of Online Victimization: School Violence and Delinquency." *Journal of School Violence* 6(3): 89–112.
65. Collier, A. (2011, July 19). A cyberbullying epidemic? No! [Web log comment]. Retrieved from <http://www.netfamilynews.org/?p=30592>.
66. Nielsen, J. (2000). Why You Only Need to Test with 5 Users. Retrieved from <http://www.useit.com/alertbox/20000319.html>.
67. Wall: How to use the Wall feature and Wall privacy. (2011). Retrieved from <http://www.facebook.com/help/?page=174851209237562>.
68. Cyberbullying. (2011). Retrieved from <http://www.kidshelp.com.au/teens/get-info/hot-topics/cyber-bullying.php>.
69. Nielsen, J. (2011). Ten Usability Heuristics. Retrieved from http://www.useit.com/papers/heuristic/heuristic_list.html.
70. Stop Cyber-Bullying. (2011). Retrieved from <http://stopcyberbullying.org>.
71. Reporting abuse or policy violations. (2011). Retrieved from <http://www.facebook.com/help/?page=178608028874393>.
72. Lieberman, H.; Dinakar, K.; Jones, B.; , "Let's Gang Up on Cyberbullying," *Computer* , vol.44, no.9, pp.93-96, Sept. 2011.
73. Ttofi, M. M., Farrington, D. P. & Baldry, A. C. (2008). Effectiveness of programmes to reduce school bullying. Swedish National Council for Crime Prevention. Retrived from http://www.bra.se/extra/measurepoint/?module_instance=4&name=Effectiveness_of_programmes_to_reduce_school_bullying_webb.pdf&url=/dynamaster/file_archive/081023/04395cbc57201c39fa6c7f78319ea2ab/Effectiveness%255fof%255fprogrammes%255fto%255freduce%255fschool%255fbullying%255fwebb.pdf