Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying (Extended Abstract)*

Karthik Dinakar, Rosalind Picard, Henry Lieberman
Massachusetts Institute of Technology
Cambridge, MA
{karthik,picard,lieber}@media.mit.edu

Abstract

We present an approach for cyberbullying detection based on state-of-the-art text classification and a common sense 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 and construct *BullySpace*, a common sense knowledge base that encodes particular knowledge about bullying situations. We then perform joint reasoning with common sense knowledge about a wide range of everyday life topics. We analyze messages using our novel *AnalogySpace* common sense 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 among 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 helps 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.

1 Introduction

Cyberbullying or harassment on social networks 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. Statistical methods such as support vector machines and Bayesian networks have been successfully deployed as spam filters. While spam messages are sent nearly identically to many people, online bullying is more personalized, varied and contextual. The structure of this paper is as follows: we show the effectiveness of traditional supervised learning methods in detecting explicit profane and negative language and also their limitations in modeling indirect and subtle forms of abuse. We then describe a model of common sense reasoning to address this limitation. We use these models to power reflective user-interaction to foster empathy among social network participants, automatically index educational material for victims and perpetrators, and a dashboard to help site moderators prioritize user complaints. Readers should consult the original journal article [**?**] for a detailed treatment of each section.

1.1 Corpora

Label	# +ve comments	# -ve comments
Sexuality	627	873
Race and Culture	841	659
Intelligence	809	691

Table 1: The YouTube dataset was annotated by 3 annotators using codes for **Sexuality** (Negative comments involving attacks on sexual minorities and sexist attacks on women); **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; **Intelligence**- Comments attacking the intelligence and mental capacities of an individual.

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 (e.g., linear algebra and photoshop tutorials); these comments were annotated for sensitive topics as shown in Table 1. Via Formspring, a popular social network for teenagers, we received a dataset of anonymized instances of bullying that were either user-flagged or caught by their moderation team. The Formspring

*This is a minor revision of the work published in ACM Transactions on Interactive Intelligent Systems, Vol. 2, No. 3, Article 18 http://dx.doi.org/10.1145/2362394.2362400..

	Naive Bayes			Rule-based JRip			Tree-based J48			SVM (poly-2 kernel)		
	Acc.	F1	kappa	Acc.	F1	kappa	Acc.	F1	Kappa	Acc.	F1	kappa
Sexuality	66%	0.67	0.657	**80%**	0.76	0.598	63%	0.57	0.573	66%	0.77	**0.79**
Race and Culture	66%	0.52	0.789	**68%**	0.55	0.789	63%	0.48	0.657	66%	0.63	**0.71**
Intelligence	72%	0.46	0.467	**70%**	0.51	0.512	70%	0.51	0.568	72%	0.58	**0.72**
Mixture	63%	0.57	0.445	63%	0.60	0.507	61%	0.58	0.456	66%	0.63	0.653

Table 2: Table showing classes of models to detect explicit bullying language pertaining to **(1) sexuality**, **(2) race & culture** and **(3) intelligence**. Binary classifiers were trained for each label using the feature space design shown in Table 3. Binary classifiers outperform their multiclass counterparts: JRip and Support Vector Machines were the best performing in terms of accuracy and kappa values. Results shown here are measured against the held-out test set.

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.

2 Modeling explicit bullying language using supervised learning

In this section, we attempt to show both the effectiveness and limitations of traditional supervised learning methods to detect cyberbullying. Since explicit verbal abuse involves the use of stereotypical slang and profanity as recurring patterns, they lend themselves nicely to traditional supervised learning methods. We hypothesize that instances of cyberbullying where the abuse is more indirect and 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 Nave Bayes a rule-based learner [**?**], a tree-based learner [**?**], and support-vector machines [**?**]. We conduct two experiments: first, training binary classification models to predict a label and second, training of multiclass classification models to predict a set of labels for a given comment.

2.1 Feature space design

We divide the feature space into general features shared across multiple labels, and label-specific features. The intuition behind the feature space design is informed by relevant research in socio-linguistics and interaction analysis pertinent to bullying. While negativity and profanity appear across many contexts of cyberbullying irrespective of the context, topics that are sensitive and personal to an individual (e.g race, cultural heritage, sexuality etc) have feature spaces of their own, as shown in Table 2 below:

2.2 Results & Error Analysis

As shown in Table 2, multi-class classifiers underperformed compared to binary classifiers. In terms of accuracy, JRip was the best, although the kappa values were best with SVM. SVMs high kappa values suggest better reliability for all labels. Nave Bayes classifiers for all labels perform much better than J48.

Feature	Type
TF-IDF	General
Ortony lexicon for negative affect	General
List of slurs & profanity	General
POS bigrams	General
Topic-specific unigrams & bigrams	Label-specific

Table 3: Main categories of features: general features shared across labels and features specific to a topic. The combination of a sensitive topic and use of slurs/profanity is indicative of harassment.

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 contain commonly used forms of abuse or profanity or expressions denoting a negative tone. For example, consider the following instance:

u1 `hey we didnt kill all of them, some are still alive today. And at least we didnt enslave them like we did the monkeys, because that would have been more humiliating`

In the instance shown above (pertaining to race) contain unigrams and expressions that lend them to be positively classified by the models. Instances such as the ones shown, 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.

u2 `they make beautiful girls, especially the one in the green top`

u3 `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. 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 associ-

ated with girls 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 how we use can use common sense reasoning to overcome these limitations with supervised learning methods.

3 Modeling subtle abuse with common sense reasoning

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 peoples goals and emotions and the objects *properties* and *relations* that can help disambiguate and contextualize language. **Open Mind Common Sense (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 recognize 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. Common sense 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 exam- ple, 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.

3.1 ConceptNet & AnaologySpace

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 males face, for instance, the two concepts are *beard* and *male*, the relation is `IsA`, and the polarity is positive. For the assertion *People dont 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.

AnalogySpace can then be constructed as follows: 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 are ordered from largest to smallest, while 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 **cosine similarity** between concepts or between features in AnalogySpace.

3.2 BullySpace

A key ingredient in tackling implicit ways of insulting another person is to transform commonly used stereotypes and social constructs into a knowledge representation. For example *"put on a wig and lipstick and be who you really are"* where 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 dont 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 employed 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 preceding example, the assertions added were as follows:

a1 `lipstick is used by girls`
a2 `lipstick is part of makeup`

```
a3 makeup is used by girls
a4 a wig is used by girls
a5 a toupee is used by men
```

We build a set of more than 200 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 AnalogySpaces 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 dimen- sionality of the matrix by selecting only the top k = 100 set of principal components. We now have the basic machinery required to perform common sense reasoning.

3.3 Results & Error Analysis

Annotators	#1	#2	#3
Disagreed	15	12	16
Agreed	35	38	34

Table 4: Expert evaluation of Formspring instances of BullySpace scores. An error anlysis discussed below, points to concept sparsity in BullySpace

We build a test set of LGBT issues 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 on 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 the detection approach that we take in this article 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 common sense reasoning model. The results are shown in Table 4.

"George Michael or Elton John?", an example where annotators disagreed, points to a sparsity problem in our knowledge base. This instance received an extremely high score for the concept boy due to the names of the individuals mentioned. However, a deeper analysis shows that the individuals are celebrity singers who also have one thing in common: 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*.

4 Reflective User Interaction

We simulate Facebook by building a pseudo version of the same for design reflective user interaction. We use the society of models created above to power such interaction. We focus on helping victims and perpetrators reflect on their behavior by using the following strategies: **(1) introduce action delays** upon detecting harassing language about to be posted, **(2) inform users of the hidden consequences of their actions** before they post abuse and **(3) suggest educational material** for victims, bystanders and perpetrators. To help moderators of the social network help prioritize user flagged messages, we create an **"air-traffic" type dashboard** designed to alert moderators on bullying outbreaks viewed from the level of the social graph. Readers should consult our original journal article for a detailed, granular treatment and evaluation of our reflective user interaction paradigms.

5 Ongoing & future work

Our work in modeling the detection of textual cyberbullying and using such detection to power reflective user interaction has led us in unexpected and exciting directions. Since the publication of this work in 2012, we have examined online adolescent distress using stacked generalization [**?**] and designed mixed-initiative models and reflective user interfaces to help distressed adolescents online with encouraging success [**?**]. This work has inspired human-in-the-loop approximate posterior inference in probabilistic graphical models, and we have used it to power interfaces and real-time topic models for online crisis counseling of adolescents [**?**]. We are currently embarking on the use of a family of latent variable models to model, understand and predict self-harm in adolescents, a phenomenon that is not very well understood in the field of abnormal psychology.

6 Acknowledgement

We wish to thank our collaborators, Dr. Bob Selman, Dr. Matthew Nock and Emily Weinstein of Harvard University, Allison Chaney of Princeton, Dr. David Blei of Columbia University, Dr. Eric Horvitz of Microsoft Research, Crisis Text Line, Boston Samaritans and MTV for their boundless enthusiasm and judicious input, and to Reid Hoffman for the Hoffman Fellowship.

References

[Abelson *et al.*, 1985] Harold Abelson, Gerald Jay Sussman, and Julie Sussman. *Structure and Interpretation of Computer Programs*. MIT Press, Cambridge, Massachusetts, 1985.

[Baumgartner *et al.*, 2001] Robert Baumgartner, Georg Gottlob, and Sergio Flesca. Visual information extraction with Lixto. In *Proceedings of the 27th International Conference on Very Large Databases*, pages 119–128, Rome, Italy, September 2001. Morgan Kaufmann.

[Brachman and Schmolze, 1985] Ronald J. Brachman and James G. Schmolze. An overview of the KL-ONE knowledge representation system. *Cognitive Science*, 9(2):171–216, April–June 1985.

[Gottlob *et al.*, 2002] Georg Gottlob, Nicola Leone, and Francesco Scarcello. Hypertree decompositions and tractable queries. *Journal of Computer and System Sciences*, 64(3):579–627, May 2002.

[Gottlob, 1992] Georg Gottlob. Complexity results for nonmonotonic logics. *Journal of Logic and Computation*, 2(3):397–425, June 1992.

[Levesque, 1984a] Hector J. Levesque. Foundations of a functional approach to knowledge representation. *Artificial Intelligence*, 23(2):155–212, July 1984.

[Levesque, 1984b] Hector J. Levesque. A logic of implicit and explicit belief. In *Proceedings of the Fourth National Conference on Artificial Intelligence*, pages 198–202, Austin, Texas, August 1984. American Association for Artificial Intelligence.

[Nebel, 2000] Bernhard Nebel. On the compilability and expressive power of propositional planning formalisms. *Journal of Artificial Intelligence Research*, 12:271–315, 2000.