Modeling the Detection of Textual Cyberbullying

Karthik Dinakar[+] *karthik@media.mit.edu*

Roi Reichart[*] *roiri@csail.mit.edu*

Henry Lieberman[+] *lieberman@media.mit.edu*

[+]MIT Media Lab, [*]Computer Science & Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139 USA

Abstract

The scourge of cyberbullying has assumed alarming proportions with an ever-increasing number of adolescents admitting to having dealt with it either as a victim or as a bystander. Anonymity and the lack of meaningful supervision in the electronic medium are two factors that have exacerbated this social menace. Comments or posts involving sensitive topics that are personal to an individual are more likely to be internalized by a victim, often resulting in tragic outcomes. We decompose the overall detection problem into detection of sensitive topics, lending itself into text classification sub-problems. We experiment with a corpus of 4500 YouTube comments, applying a range of binary and multiclass classifiers. We find that binary classifiers for individual labels outperform multiclass classifiers. Our findings show that the detection of textual cyberbullying can be tackled by building individual topic-sensitive classifiers.

Introduction

That cyberbullying has grown as a social menace, afflicting children and young adults is well known. Not limited to children and young adults, cyberbullying has also increased in the workplace [1]. According to recent studies, almost 43% of teens in the United States alone reported being victims of cyberbullying at some point in time [2].

Cyberbullying, like traditional forms of bullying, has a deeply negative impact on the victim, especially children and young adults in their formative years. The American Academy of Child and Adolescent Psychiatry says that victims of cyberbullying often experience significant emotional and psychological suffering [3]. Many of these cases have tragically ended in suicides, underlining the grave nature of this problem.

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' [4]. Social scientists such as Danah Boyd have described four aspects of the web that changes the very dynamics of bullying and magnifies it to new levels: persistence, searchability, replicability and invisible audiences [5].

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.

Previous work related to cyberbullying has centered on extensive surveys unearthing the scope of the problem and on its psychological effects on victims. Little attention if any, has been devoted to its detection, beyond regular-expression-driven systems based on keywords. There has been a dearth of computationally driven approaches for its detection.

In this paper, we focus on the detection of textual cyberbullying, which is one of the main forms of cyberbullying**.** We use a corpus of comments from YouTube videos involving sensitive topics related to race & culture, sexuality and intelligence i.e., topics involving aspects that people cannot change about themselves and hence become both personal and sensitive. We pre-process the data, subjecting it to standard operations of removal of stop words and stemming, before annotating it to assigning respective labels to each comment.

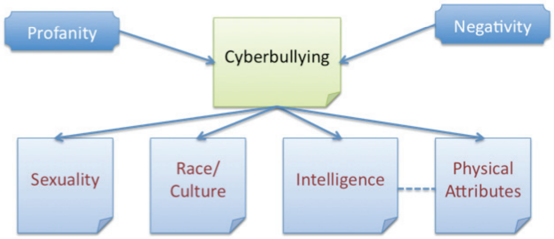


Figure 1: Problem Decomposition: A comment involving a combination of negativity or profanity and topics that are personal and sensitive are those can are most hurtful, forming candidates for cyberbullying.

We decompose the detection of bullying into subproblems involving text classification. We perform two experiments: a) training binary classifiers to ascertain if an instance can be classified into a sensitive topic or not and b) multiclass classifiers to classify an instance from a set of sensitive topics. Our findings show that individual classifiers that classify a given comment into a specific label or not fare much better than multiclass classifiers involving a set of labels.

Problem Decomposition

Within social networking websites, it is a common practice for people to make comments and post messages.

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 becomes a worthy candidate that may qualify as cyberbullying.

In addition, if the same comment also has a negative connotation and contains profane words, the combination of rudeness and talking of sensitive, personal topics can be extremely hurtful.

For most children in middle school and young adults, the sensitive list of topics often assume one of the following: physical appearance, sexuality, race & culture and intelligence. When a comment or a message on a social networking website involving these sensitive topics are made with rudeness and profanity, and in front of the victim's network of friends, it can be very hurtful. In fact, 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.

The problem of detecting hurtful messages on social networking sites can viewed as the following: classifying messages as speaking on sensitive topics and detecting negativity and profanity. The problem then lends itself into a bag-of-words driven text classification experiment.

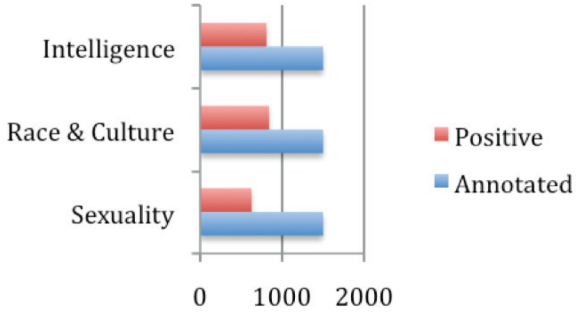


Figure 2: Corpus: Comments downloaded were annotated and grouped into three categories of 1500 instances each under sexuality, race & culture and intelligence. 627, 841 and 809 instances were found to be positive for sexuality, race & culture and intelligence respectively.

Granularity in the arc of cyberbullying

Social scientists investigating this menace describe the goals of cyber bullies to harm, disrepute or embarrass a victim through 'repeated' acts such as the posting of inappropriate text messages [6]. As such, the arc of textual cyberbullying consists of a sequence of messages targeting a victim by a lone perpetrator or by a group of individuals.

Exploiting this level of granularity by detecting individual messages that might eventually lead to a tragic outcome assumes importance for two reasons: the design of pre-emptive and reactive intervention mechanisms. In this work, we focus on detecting such individual messages.

Corpus

The dataset for this study was obtained by scraping the social networking site www.youtube.com for comments posted on videos. 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.

Most comments on YouTube can be described as standalone, with users expressing opinions about the subject and content of the video. While some of the comments were made as responses to previously posted ones, there were no clear patterns of dialogue in the corpus. As such, we therefore treat each comment as stand-alone, with no conversational features.

Using the YouTube PHP API, we scraped roughly a thousand comments from controversial videos surrounding sexuality, race & culture and intelligence. We were constrained by the limitations posed by YouTube of being able to download an upper limit of up to a 1000 comments per video. The total number of comments downloaded overall greater than 50,000.

The downloaded comments were grouped into clusters of physical appearance, sexuality, race &culture and intelligence. 1500 comments from each cluster were then hand annotated to make sure that they had the right labels assigned to them. Those comments that were not related to the cluster (for e.g., the comment 'Lol, I think that is so funny') were given a neutral label 'none'. Each cluster had few comments that belonged other labels too.

Data preprocessing

Each dataset was subjected to three operations: removal of stop-words, stemming and removal of unimportant sequence of characters. Sequences of characters such as '@someuser','lollllll','hahahahaha', etc., were expunged from the datasets.

Annotation

The comments downloaded from all the videos were arranged in a randomized order prior to annotation. Two 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:

Sexuality: Negative comments involving attacks on sexual minorities and sexist attacks on women.

Race and Culture: Attacks bordering on racial minorities (eg African-American, Hispanic and Asian) and cultures (eg Jewish, Catholic and Asian traditions) including unacceptable descriptions pertaining to race and stereotypical mocking of cultural traditions.

Intelligence: Comments attacking the intelligence and mental capacities of an individual.

Inter-rater agreement

Annotated comments with an inter-rater agreement of kappa >= 0.4 were selected and grouped under each label until 1500 comments were available for each of them.

Experiment Setup

The datasets for each cluster were divided into 50% training, 30% validation and 20% test data. Each dataset was subjected to data preprocessing to clean and massage the data. The next task was to select and populate the feature space for three supervised learning methods along with a Naïve Bayes classifier.

a) Repeated Incremental Pruning to Produce Error Reduction, more commonly known as JRip, is a propositional rule learner proposed by Cohen et.al [7]. It is a two-step process to incrementally learn rules (grow and prune) and then optimize them.

b) J48 is a popular decision tree based classifier based on the C4.5 method proposed by Ross Quinlan [8]. It uses the property of information gain or entropy to build and spilt nodes of the decision tree to best represent the training data and the feature vector.

c) Support-vector machines [9] 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 cite support-vector machines as the state of the art methods for textual classification. We use a poly-2 kernel to train our classifiers [10].

In the first experiment, binary classifiers using the above were trained on each of the three datasets for each of the labels, namely, 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.

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.

The intuition behind this is as follows: negativity and or profanity is general across all 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).

General features

The general features consists of TF-IDF weighted unigrams, the Ortony lexicon of words denoting negative connation, a list of profane words and frequently occurring POS bigram tags observed in the training set across each of the datasets.

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 words that were observed in the training data .

TF-IDF

The TF-IDF (term frequency-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. Given comments c_1,c_2,…c_j, where a comment c containing words w_1,w_2,…,w_k, the word frequency relative to a comment and its inverse comment frequency is a simple calculation as follows:

Ortony Lexicon for negative affect

The Ortony lexicon [11] (containing a list of words in English that denotes the affect) was stripped off 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.

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.

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.

Evaluation

The aforementioned models 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 as shown in the tables 4 and 5.

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.

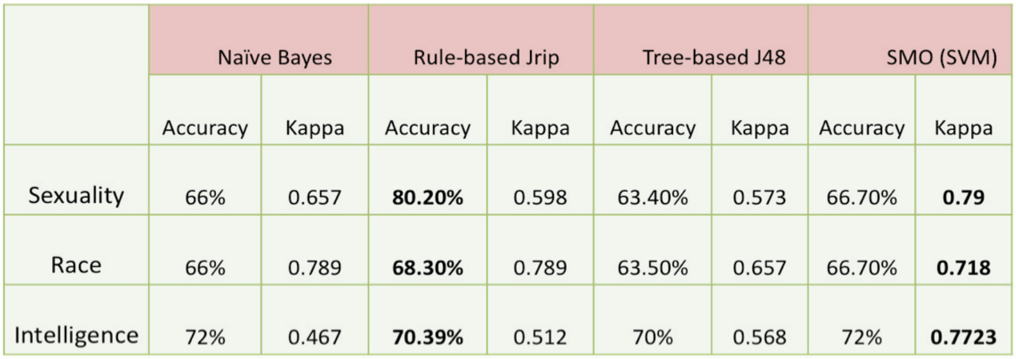
Prior work on the assessment of classifiers suggests that accuracy alone is an insufficient metric to gauge reliability. The kappa statistic (Cohen's kappa), which takes into account agreement by chance, has been argued as a more reliable metric in conjunction with accuracy [12]. We evaluate each classifier in terms of both the accuracy 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 lesser compared to SVM. SVM's high kappa values suggest better reliability for all labels. Naïve Bayes classifiers for all labels perform much better than J48.

The results suggest the building of systems with topic-specific models for classifying posts involving sensitive topics. Instances that are positively classified by such models are very likely to be candidates that pass for cyberbullying.

Error Analysis

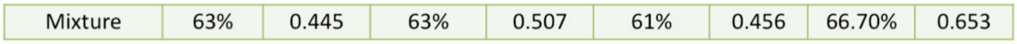
An error analysis on the results reveals that instances where bullying is apparent and blatant is simple to model. Such instances either contains commonly used forms of abuse or profanity or expressions connoting negativity. For example, consider the following instances in figure 6.



	Naïve Bayes		Rule-based Jrip		Tree-based J48		SMO (SVM)	
	Accuracy	Kappa	Accuracy	Kappa	Accuracy	Kappa	Accuracy	Kappa
Sexuality	66%	0.657	**80.20%**	0.598	63.40%	0.573	66.70%	**0.79**
Race	66%	0.789	**68.30%**	0.789	63.50%	0.657	66.70%	**0.718**
Intelligence	72%	0.467	**70.39%**	0.512	70%	0.568	72%	**0.7723**

Figure 4: Binary classifiers for individual labels

- Binary classifiers trained for individual labels fare much 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.



Mixture	63%	0.445	63%	0.507	61%	0.456	66.70%	0.653

Figure 5: Multiclass classifiers for the merged dataset

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 atleast we didn't enslave them like we did the monkeys, because that would have been more humiliating*

Figure 6: Two instances with patterns of words and phrases that is simple to model. The first instance is from a video on gay marriage, while the second is from a video on the life of Martin Luther King.

u3 most of them come north and are good at just mowing lawns

u4 you're intelligence is so breathtaking!!!!!!

u5 she will be good at pressing my shirt

Figure 7: The first comment was made in reference to an immigrant community, while the second was for a video involving a beauty pageant contest, where the contestant's answer to a question was widely viewed as less than satisfactory. The third was for a video about a famous female politician.

Both the instances shown above (the first pertaining to sexuality and the second pertaining to race) contain words and expressions that lend them to be positively classified by the models. What is more difficult to model is sarcasm, such as the following two instances in figure 7.

Instances employing sarcasm were frequently misclassified, especially in the absence of a contextually relevant word, profanity or negativity. In the first instance, the lack of profanity or negativity likely mislead the classifier into assigning it the label 'none'. In the second instance, the lack of any contextually relevant words led to its misclassification, even though it was annotated as sexist.

Related Work

Much of the work related to cyberbullying as a phenomenon is in the realm of social sciences and psychiatry. As such, this problem has not been attacked from the perspective of statistical models for detection and intervention.

The related work to computational ways of detecting cyberbullying can therefore be seen from three angles – the social sciences & psychiatry, text classification & information extraction, and their application to similar kinds of real world problems.

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 [13].

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 [14].

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.

Text categorization, Topic Modeling & Information Extraction

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. We use a nonlinear poly-2 kernel version of support vector machines as one of our methods [15].

Probabilistic models of unearthing semantic content from documents through the extraction of latent topics are active areas of research in natural language processing.

Topic modeling techniques from Latent Dirichlet Allocation [16] (LDA) to Hidden Topic Markov Models (HTMM) [17] have been used in applications pertaining to information retrieval. Named-entity recognition and terminology extraction techniques have also been used in a variety of applications.

Similar real-world applications

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 [18].

Support vector machines have been shown to be effective in detecting email spam [19], while creative feature space designs for machine learning approaches to modeling the detection of vandalism have been shown to be effective.

Future Work

In this work, we treat each comment on its own and don't consider other aspects to the problem as such the pragmatics of dialogue and conversation and the social networking graph. Taking into account such features will likely prove more useful on social networking websites such as Facebook or Formpsring.

More interestingly, it would be interesting to apply a semi-supervised learning techniques to leap over the problem of hand- annotation of data, which is potentially intractable given the huge corpus of data that is available across multiple social networking websites.

Sentiment mixture models that account for multiple views of a given post, as well topic-author-community models that consider social interaction variables would be interesting to pursue.

Reflective user interaction

Detection of the granular aspects of cyberbullying would only likely be the first step. The question then becomes what is done after detection.

User-interfaces need to adapt in subtle, yet effective ways to help assuage the problem of cyberbullying in the context of social networks. Prior work in goal-oriented interfaces has focused on facilitating the user experience to accomplish task or goals, but not to prompt reflection on their intent or consequences of their actions.

New user-interface and experience design patterns are needed to promote end-user reflection and to halt the progression of passive textual abuse in online discussion and social interaction conversations.

Conclusion

In this paper, we focus on the problem of detecting textual cyberbullying in stand-alone posts with a dataset of YouTube comments. We decompose the problem into detection of topics that are of a sensitive and personal nature. Labels are of a personal nature and instances that have a negative connotation and might include profanity are likely to be an instance of cyberbullying.

Our experiments show that building label-specific classifiers are more effective than multiclass classifiers at detecting such sensitive messages. Our analysis shows that blatant bullying involving patterns of verbal abuse and profanity are easier to model than expressions involving sarcasm and euphemism.

This work can be extended to include the pragmatics of dialogue and the social networking graph for better modeling of the problem. Detection of textual cyberbullying is the first step; computational intervention mechanisms for reflective user interfaces would be a logical next step.

Acknowledgement

We wish to thank Professor Rosalind Picard and Birago Jones from the MIT Media Lab for their advice and guidance. We also thank Danah Boyd from Microsoft Research for her insights into the world of teenagers and their experience with the social media. We also thank the students from the Software Agents Group at the MIT Media Lab for their inputs.

References

1. 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 doi:10.1177/1461444809341263

2. Ybarra, M. (2010, February 25). *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: Forging a Shared Agenda. Houston, TX.

3. Facts for families, the American Academy of Child Adolescent Psychiatry, available at http://www.aacap.org/galleries/ FactsForFamilies/80_bullying.pdf.

4. Cyberbullying, The National Crime Prevention, available at http://www.ncpc.org/cyberbullying

5. Boyd, Danah. (2007) "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.

6. Lenhart, A. Cyberbullying 2010: What the research tells us. Washington, DC: Pew Youth Online Safety Group, 2010

7. Cohen,W.W., & Singer, Y. (1999). A simple, fast, and effective rule learner. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*

8. Ross Quinlan (1993). *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, CA.

9. Corinna Cortes and V. Vapnik, "Support-Vector Networks", Machine Learning, 20, 1995. Available at http://www.springerlink.com/content/k238jx04hm87j80g/

10. 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

11. Andrew Ortony, Gerald L. Clore, Mark A. Foss: The Referential Structure of the Affective Lexicon. Cognitive Science 11(3): 341-364 (1987)

12. Carletta, Jean. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics,* 22(2):249-254.

13. Qing Li. 2007. New bottle but old wine: A research of cyberbullying in schools. *Comput. Hum. Behav.* 23, 4 (July 2007), 1777-1791. DOI=10.1016/j.chb.2005.10.005 http://dx.doi.org/10.1016/j.chb.2005.10.005

14. Smith, P. K., Mahdavi, J., Carvalho, M., Fisher, S., Russell, S., & Tippett, N. (2008). Cyberbullying: Its nature and impact in secondary school pupils. Journal of Child Psychology & Psychiatry, 49, 376−385.

15. Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM Comput. Surv.* 34, 1 (March 2002), 1-47. DOI=10.1145/505282.505283 http://doi.acm.org/10.1145/505282.505283

16. D. Blei and J. Lafferty. Topic Models. In A. Srivastava and M. Sahami, editors, *Text Mining: Theory and Applications*. Taylor and Francis, 2009.

17. A. Gruber, Y. Weiss, and M. Rosen-Zvi. Hidden Topic Markov Models. In Proc. of the Conference on Artificial Intelligence and Statistics, 2007

18. K. Smets, B. Goethals, and B. Verdonk. Automatic Vandalism Detection in Wikipedia: Towards a Machine Learning Approach. In Proc. of WikiAI at AAAI'08

19. D. Sculley and Gabriel M. Wachman. 2007. Relaxed online SVMs for spam filtering. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '07). ACM, New York, NY, USA, 415-422. DOI=10.1145/1277741.1277813 http://doi.acm.org/10.1145/1277741.1277813

20. Reid Priedhorsky, Jilin Chen, Shyong (Tony) K. Lam, Katherine Panciera, Loren Terveen, and John Riedl. 2007. Creating, destroying, and restoring value in wikipedia. In *Proceedings of the 2007 international ACM conference on Supporting group work* (GROUP '07). ACM, New York, NY, USA, 259-268. DOI=10.1145/1316624.1316663 http://doi.acm.org/10.1145/1316624.1316663