GabGuard: Automatically Detecting Bullying, Depression, and Victimization in Children's Social Networks

Kiley Sobel
Sketch Recognition Lab
Department of Computer Science and Engineering
Texas A&M University
ksobel@g.hmc.edu

Stephanie Valentine
Sketch Recognition Lab
Department of Computer Science and Engineering
Texas A&M University
valentine@cs.tamu.edu

Tracy Hammond
Sketch Recognition Lab
Department of Computer Science and Engineering
Texas A&M University
hammond@cs.hmc.edu

ABSTRACT
Online bullying has become a major issue in our society as more kids enter into the world of social networking and the Internet at young ages. Now more than ever it is important to teach children how to use the Internet safely and appropriately to ensure that they will be good cyber citizens once they reach adulthood. We are developing Kidgab, a school-based social network for students under thirteen to learn how to be responsible digital citizens in a monitored, supportive environment. Automated sentiment analysis is one method to detect negative attitudes (such as depression, victimization, abuse, etc.) and cyberbullying in the network for teachers to review. We developed a system called GabGuard to detect negative sentiment in messages and status updates in Kidgab using a spell-checker, a dictionary of slang words and phrases, and natural language processing applications like word spotting and semantic polarity classification. Testing the detection on a corpus of children's forum posts indicated that the system needs improvement in relation to context-specific messages. This work provides the foundation for more precise, in-depth cyberbullying and negative sentiment detection in Kidgab.

Author Keywords
cyberbullying detection; sentiment analysis;
children-computer interaction; social networking.

ACM Classification Keywords
H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

INTRODUCTION
Cyberbullying has become a major issue in our technological society as children enter into the world of social networking and the Internet at young ages [11] [13]. The emergence of depression, victimization, and faceless bullying online has manifested our need to teach children how to use the Internet appropriately and safely in order to produce a future community of respectful and responsible digital citizens [19].

There has been a limited amount of research in the area of cyberbullying detection [7] [8] [9] [10], especially in regards to the domain of kids under the age of thirteen. This research is extremely recent and exploratory, using learning-based algorithms and statistical analyses with Naive Bayes classifiers and WEKA, a suite of machine learning software.

Kidgab
We are developing Kidgab, a school-based social network for nine to thirteen year olds to learn how to be good cyber citizens in a monitored, supportive environment. The network will engage students, parents, and teachers to create a space for users to interact educationally and socially. Automatic negative sentiment detection, specifically related to cyberbullying, is necessary in the network. It will allow teachers to monitor the children's behavior by alerting the site administrators of online activity that needs to be dealt with in person.

Pre-Teen Netspeak and Sentiment
Past research in negative sentiment analysis focuses on databases of text written in standard language. For example, published texts compose the corpus [18] or researchers remove words from the text corpus (e.g. 'lollllllll' and 'hahahahaha') that may actually have impact on the sentiment [10].

Sood, Antin, and Churchill explored the use of profanity online and how users censor profanity with slang or by replacing letters with punctuation [17]. They found that list-based approaches of profanity detection are insufficient due to their failure to adapt to constantly evolving language, misspellings, and censor methods. However, using Turney's work, past researchers have used Urban Dictionary to build a sentiment dictionary and determine semantic orientation scores [12].

Internet users employ different orthographic devices to decrease time entering text, display emotional affect, or simulate spoken language [16]. These include vowel omission, letter and number homophones, acronyms, shortening words, accent simulation, reduplication of sounds/repetition of letters, emoticons, text pictures, and other features [16]. Additionally, pre-teens do not have sophisticated or perfect styles of speech, which may be difficult to decipher without learning-based algorithms.

Because pre-teen netspeak is complex, continually progressing, and being augmented, analyzing this online language is a hard problem.

METHOD

We developed GabGuard, a system that automatically detects cyberbullying, depression, victimization, and general negative sentiment in children's messages. GabGuard determines if a message is negative by scoring each word in the message using SentiWordNet [6], an opinion mining extension of the lexical database WordNet [15]. The system then calculates statistical information about the negativity in the message, according to Ohana [14]: the number of negative words in a message, the strength of the negativity (average negativity score over negative words), and the percentage negativity in the message (the number of negative words over the total number of words).

Before the system gives a word a negativity score, GabGuard preprocesses the word in order to accurately determine its semantic polarity.

Part of Speech Labeling

GabGuard uses the Natural Language Toolkit (NLTK), a Python suite for natural language processing [1], to classify the part of speech of each word in the message. SentiWordNet can use the part of speech to more accurately get the sentiment polarity of the word (particularly when there are multiple forms and/or definitions of the word).

Simplification

The system eliminates any non-alphabetic characters from a word and puts the word into all lower-case letters.

Spell-checking

Using work by Novrig [2], GabGuard performs spell-check on the word. Using basic probability theory [2], the spell-checker is trained on a large text file provided by Novrig to which we added slang [5] and swearwords [4]. The program then uses the concept of edit distance (the number of edits- deletions, transpositions, alterations, or insertions- it takes to turn one word into another) to perform the spell-check. It checks for correct spellings up to two edits away from the original word [2].

Because people often use different orthographic devices in online text [16], we decided to focus on the replication of letters to improve word detection. GabGuard attempts to find the correct spelling of words not recognized by the dictionary by removing duplicate letters and rechecking the spelling. If a word has numerous repeated letters, a spell-checker based on edit distance will not fix the spelling because the word is too many edits away from the correct spelling. Thus, with duplicate letters removed, the word should theoretically be closer in edit distance away from the true spelling.

Swearword Spotting

GabGuard holds a dictionary of swearwords from the Swear Word List, Dictionary, Filter, and API [4]. GabGuard does not tolerate any swearwords in the message.

Finding Synonyms of Slang Words and Phrases

By scraping the website Urban Dictionary [5] of the most popular entries, we obtained a dictionary of 3,540 slang words, acronyms, and phrases with their definitions and lists of their synonyms.

When a term does not exist in the SentiWordNet dictionary, GabGuard checks to see if the term exists in the dictionary of Urban Dictionary terms. If the term is slang, a function returns the first non-slang entry in the list of synonyms for the slang term. The system uses this new non-slang word to determine the original word's negative polarity.

Determining Negative Polarity Score

GabGuard assigns each word a negativity score by using SentiWordNet positive polarity scores and negative polarity scores for words in WordNet. WordNet groups words into sets of synonyms called synsets. A word may have multiple synsets due to how words can be used differently in sentences. For example, "value" can be a noun or a verb; the noun "value" has different meanings (worth or beliefs) while the verb "value" has different meanings (thinking highly of or appraising). In addition, the synsets can have different sets of polarity scores. For example, the noun "dog" (the animal) has a neutral score, and the verb "dog" (to plague) has a negative score.

Similar to the work of Ohana [14], GabGuard assigns the polarity of a word by setting a minimum threshold difference between positive and negative scores for each synset and a minimum threshold difference between the average positive and negative polarity scores over all the synsets. If at least one synset has a difference that exceeds a threshold and the average difference also exceeds a threshold, then the word can be classified as negative when the negative score(s) outweigh the positive score(s).

Figure 1. How GabGuard analyzes the word "depressed" written with duplicated letters.

ALGORITHM OVERVIEW

First, the system gets the part of speech of each word in the message (which SentiWordNet will use). Then for each word, GabGuard:

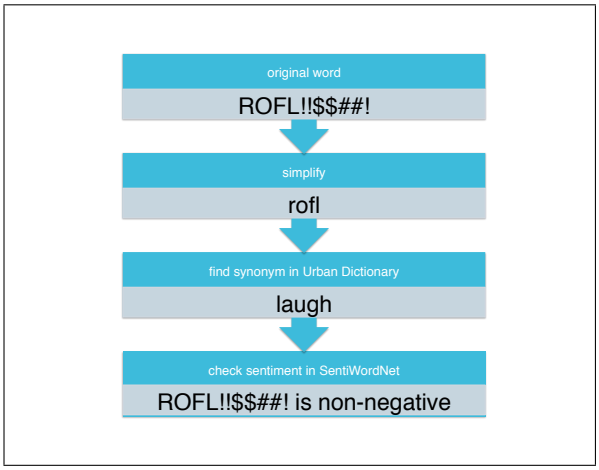

Figure 2. How GabGuard analyzes the acronym/slang term "ROFL" (Rolling On The Floor Laughing).

1. Removes non-alphabetic characters and makes the word lowercase.
2. Runs spell-check on the word.
3. If the new spelling of the word is the same as the original word (which indicates the word may be spelled correctly, be slang, or be too many edit distance units away from the correctly spelled word due to letter repetition):
    - Removes duplicates and then re-runs spell-check on the word.
4. Checks if the word (original or new spelling) is a swear-word. If so, returns the word as negative.
5. If the original word is in the SentiWordNet dictionary, returns the sentiment from SentiWordNet.
6. If the original word is in the Urban Dictionary, returns the sentiment of the first synonym that is in SentiWordNet.
7. Otherwise, the word may be misspelled; returns the sentiment of the new spelling of the word.

Figure 1 and Figure 2 show how different words transform through the sentiment classification algorithm.

TESTING

We scraped Kidzworld [3], a website established in 2001 for children and teens ages nine to fifteen. At the time we scraped the website for forum posts, there were 51 active members and 77 guests on the site, 151,315 threads, 3,373,614 posts, and 1,892,732 members. We obtained a total of 7,033 posts from the first 40 pages of the general forum and classified the first 3,000 of the messages as negative or non-negative in sentiment.

Figure 3 shows an example of victimization in the forum, and Figure 4 shows an example of cyberbullying in the forum. We considered these types of messages negative in sentiment. We ran GabGuard on these 3,000 forum posts to compare to the human classification. If GabGuard found any negative sentiment in the message, it classified the message as negative.

> last week i got called fat and people were making fun of the way i look . today people were saying i've got eveil eyes and look evil don't wanna tell anyone ,teachers don't do anything only told 3 friends i should tell some more but what should i do p.s please help asap

Figure 3. Example victimization in Kidzworld general forum.

> Dont you dare start a roleplay! everytime u guys come in u start a gay as hell sh1t im tired of it if ur to comnent and chat then go do so but dont roleplay u guys took over last one 1 to post wins and made it boring i like to rp but u guys just ruin every forum i enjoy messege to orange sherbert ....im having a bad day dont get on my bad side

Figure 4. Example of cyberbullying in Kidzworld general forum.

RESULTS

Overall, GabGuard significantly over-detects negativity in messages. Figure 5 displays the number of messages classified as negative and non-negative by GabGuard and by human analysis. There were a total of 1605/3000 messages (53.55%) classified equivalently between human and computer. Comparing computer analysis to the human analysis as the baseline, GabGuard detected 568/649 (87.5%) negative messages correctly and 1037/2351 (44.1%) non-negative messages correctly. Thus, there were 1314 messages falsely identified as negative in sentiment by GabGuard and 81 messages falsely identified as non-negative in sentiment by GabGuard, compared to human classification.

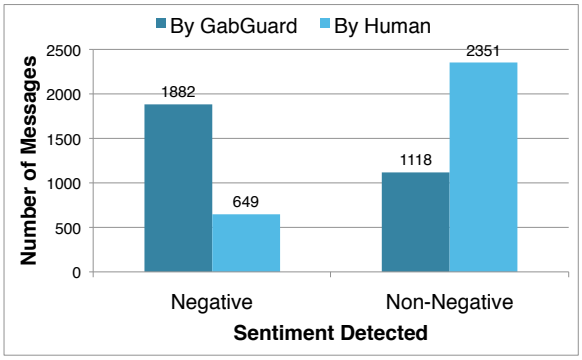

Figure 5. Sentiment detection in Kidzworld forum posts.

DISCUSSION

GabGuard's inaccuracies in negative sentiment detection are mostly due to context. The system has no way of determining if the negative words are directed at a person or at an object. Thus, messages indicating that, for instance, a movie was bad may be detected as negative, while the negative sentiment is not relevant in our case.

Homographs are also confusing to the system. The SentiWordNet step in the algorithm assigns negative polarity to a word only if the word is negative on average over all the synsets. However, this does not mean that a word with one

negative meaning is never used in a negative sense. If an Internet user called another user a "cow," our algorithm will not assign "cow" a negative score when "cow" is negative in this form.

Furthermore, the system classifies a message as negative if there are any negative words in the message, regardless of the statistics about negativity it calculated. Specifically, it does not take into account the number of negative words, the strength of the negativity, or the percent negativity to classify a message as negative. Any negativity in the message makes the entire message negative. Therefore, because one negative word in a message does not imply the entire message is negative, this results in many false positives (falsely identifying the message as negative).

FUTURE WORK

In the future, we want to improve detection in a number of different ways. First, having a corpus of data from Kidgab (or another similar children's social networking site) with messages human-labeled as negative or non-negative, ideally by authoritative figures (i.e. teachers, principals, etc.) from elementary and middle schools, would greatly simplify our problem. With this particular labeled corpus, we would have the ability to employ a more complex learning-based algorithm on the data, as opposed to only using SentiWordNet. Moreover, having these labeled messages would allow for more significant and revealing test results. Therefore, our goal is to obtain a better corpus of data and have a larger, more meaningful group label the messages (perhaps using Turkers from Amazon Mechanical Turk).

In addition, we intend to preprocess the file of Urban Dictionary terms in order to add in the sentiment of each term. This way, the sentiments of slang will already be in the dictionary of Urban Dictionary terms. This will speed up the running time for fetching slang sentiment scores.

Next, we plan to explore other orthographic devices used to express emotion (e.g. capitalization, punctuation, emoticons, etc. and perform more thorough analyses on the statistical information about the negativity scores of words in messages. Using information about the users (in social networks where that information is available) could also improve detection, as well as using more natural language processing techniques related to grammar and parts of speech.

CONCLUSION

This work provides the foundation for more precise, in-depth cyberbullying, depression, and victimization detection in children's social networks, specifically in relation to Kidgab.

GabGuard is the first step in discovering how we can detect negative sentiment in children's social networks in order to decrease cyberbullying and create better digital citizens for the future.

ACKNOWLEDGMENTS

We thank the Department of Computer Science and Engineering at Texas A&M University and the Committee on the Status of Women in Computing Research (CRA-W) Distributed Research Experiences for Undergraduates (DREU) program.

REFERENCES

1. `http://nltk.sourceforge.net`.
2. `http://norvig.com/spell-correct.html`.
3. `http://www.kidzworld.com`.
4. `http://www.noswearing.com`.
5. `http://www.urbandictionary.com`.
6. BACCIANELLA, S., ESULI, A., AND SEBASTIANI, F. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of LREC-10, 7th Conference on Language Resources and Evaluation* (2010), pp. 2200–2204.
7. DADVAR, M., AND DE JONG, F. Cyberbullying detection: a step toward a safer internet yard. In *Proceedings of the 21st international conference companion on World Wide Web* (New York, NY, USA, 2012), WWW '12 Companion, ACM, pp. 121–126.
8. DADVAR, M., DE JONG, F., ORDELMAN, R., AND TRIESCHNIGG, D. Improved cyberbullying detection using gender information. In *Proceedings of the 12th-Dutch-Belgian Information Workshop (DIR2012)* (2012).
9. DINAKAR, K., HAVASI, C., JONES, B., LIEBERMAN, H., AND PICARD, R. Timeout: Commonsense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Intelligent Interactive Systems (ACM TiiS) 2*, 3 (2012).
10. DINAKAR, K., REICHART, R., AND LIEBERMAN, H. Modeling the detection of textual cyberbullying. In *Proceedings of International AAAI Conference on Weblogs and Social Media, Workshop "Social Mobile Web"* (2011).
11. FOUNDATION, H. What is cyber bullying? 2011.
12. GRUHL, D., NAGARAJAN, M., PIEPER, J., ROBSON, C., AND SHETH, A. Multimodal social intelligence in real-time dashboard system. *The VLDB Journal*, 19 (2010), 825–848.
13. LIEBERMAN, H., DINAKAR, K., AND JONES, B. Let's gang up on cyberbullying. *IEEE Computer Magazine 44*, 9 (September 2011).
14. OHANA, B. Opinion mining with the sentiwordnet lexical resource. Master's thesis, Dublin Institute of Technology, 2009.
15. PRINCETON UNIVERSITY. About Wordnet. WordNet, 2010. `http://wordnet.princeton.edu`.
16. SHORTIS, T. *The Texture of Internet*. Cambridge Scholars Publishing, 2007, ch. Revoicing Txt: Spelling, Vernacular Orthography and 'Unregimented Writing'.

17. SOOD, S., ANTIN, J., AND CHURCHILL, E. Profanity use in online communities. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems* (New York, NY, USA, 2012), CHI '12, ACM, pp. 1481–1490.

18. SOOD, S. O., CHURCHILL, E. F., AND ANTIN, J. Automatic identification of personal insults on social news sites. *Journal of the American Society for Information Science and Technology 63*, 2 (2012), 270–285.

19. WINN, M. R. Promote digital citizenship through school-based social networking. *Learning and Leading Magazine* (December 2011/January 2012), 10–13.