Aggressive text detection for cyberbullying

Laura P. Del Bosque Sara E. Garza

Abstract

Aggressive text detection in social networks allows to identify offenses and misbehavior, and leverages tasks such as cyberbullying detection. We propose to automatically map a document with an aggressiveness score (thus treating aggressive text detection as a regression problem) and explore different approaches for this purpose. These include lexicon-based, supervised, fuzzy, and statistical approaches. We test the different methods over a dataset extracted from Twitter and compare them against human evaluation. Our results favor approaches that consider several features (particularly the presence of swear or profane words).

1 Introduction

The way in which people communicate has changed and evolved during the last decades [12]. Even though technology offers several benefits for young people (12-25 years old), it has also several negative effects [4]; for instance, e-mail, texting, chats, smart phones, web cams, and web sites might be used to hurt other people [3]. In fact, the continuous intentional aggression over an indefense victim via electronic media is known as *cyberbullying* [7]. For several reasons (e.g. allowing people to hide behind an alias), this kind of virtual stalking is actually more pernicious than traditional bullying [12]. Unfortunately, phenomena such as these could ultimately end in violence and suicide. Needless to say, it is important to address this problem—for example, by using information technologies to identify cases of cyberbullying.

We believe that the first step towards cyberbullying automatic identification concerns *aggressive text detection*, where we consider as aggressive any text or document that intends to offend a person or group of persons. To tackle this issue, we define a simple aggressiveness scale (0-10, where 10 is strongly aggressive) and propose several methods to score a document in terms of aggression; these methods have been selected and designed by assuming that aggressive text detection is a sub-task of *sentiment analysis*, and thus include lexicon-based, supervised, fuzzy, and statistical approaches.

To evaluate the proposed approaches, we extracted two comment datasets from Twitter, a popular microblog where users are free to express their opinions and address other users (presumably without censorship). We compared the automatically-generated scores with the manual scores from a group of evaluators. Our results, in general, show that several of the approaches are feasible, particularly those that combine different features.

The remainder of this document is organized as follows: Section 2 provides a brief background on sentiment analysis, and Section 3 presents related work. Section 4 introduces the different approaches that were employed for aggressive text scoring, and Section 5 describes experiments and results. Finally, Section 6 presents conclusions and future work.

2 Sentiment analysis

Sentiment analysis studies subjective expressions (reviews, comments, views, emotions, etc.) that are usually found on media such as blogs, discussion boards, and news [9]. This discipline is inherently complex and involves an assortment of other disciplines, such as NLP, text mining, NER, and machine learning. Sentiment analysis includes several tasks:

Document sentiment classification.- Consists of determining whether a document is positive or negative. This is also known as *polarity detection*.

Aspect-based analysis.- Consists of detecting which specific aspect is being liked or disliked.

Opinion lexicon generation.- Consists of collecting words or phrases that express sentiment.

Comparative opinion mining.- Consists of analyzing opinions that compare items or aspects.

With respect to sentiment classification, there are two main forms to fulfill this task. One of these concerns *lexicon-based methods* (also referred to as unsupervised or semantic approaches) and the other concerns *supervised learning methods*; both of these exhibit pros and cons. Lexicon-based methods usually involve searching for the document's words in a given lexicon (vocabulary) and retrieving their polarity; the document's polarity is generally determined with a *term counting* strategy [8], i.e. a strategy in which a document is classified as positive when there are more positive than negative words and vice-versa. Supervised approaches, on the other hand, learn a model that predicts the document's polarity given a set of training examples; common supervised methods for classification include naïve Bayes, neural networks, and support vector machines (SVM).

3 Related work

State-of-the-art methods for aggressive text detection within the framework of cyberbullying are oriented towards binary text classification using supervised approaches; these works—which also tend to explore different alternatives—are strongly committed towards finding an adequate set of features for performing the classification. The work by Dinakar et al. [5], for example, considers that a hurtful comment (document) covers *sensitive topics*, such as physical appearance, sexuality, race and culture, and intelligence. Their approach, consequently, trains both binary and multi-class classifiers to detect comments exhibiting these topics (separate binary classifiers are trained to decide whether a comment covers or does not cover one of the sensitive topics); the features they take into account are varied and include tf-idf unigrams (i.e., text frequency-inverse document frequency with single words), the presence of swear words (obtained from a lexicon), frequent POS bigrams (i.e., part-of-speech tag pairs) in hurtful messages, and topic-specific unigrams and bigrams. The approach is tested using JRip, J48, SVM, and naïve Bayes over a set of `Youtube` comments; results are compared against a manual classification. The JRip binary classifier was the best.

A similar approach is followed by Dadvar et al. [4]; these authors propose to consider gender information for aggressive document detection; for this reason, they train two separate classifiers (one per gender). Their features include second-person pronouns, swear words (take the most frequent also by gender), and tf-idf values. This approach is tested using an SVM to classify `MySpace` posts; their results (also compared vs. a manual annotation) show that taking gender into account, in fact, does increase precision.

Another outstanding work is the one by Nahar et al. [10]; this work extracts *semantic features* using Latent Dirichlet Allocation (LDA), and utilizes the lexicon of `noswearing.com`, tf-idf values, and second-person pronouns also as features for training an SVM. They test their approach over a dataset provided by the workshop of Content Analysis for the Web, which comprises comments from `Twitter`, `Slashdot`, and `MySpace`.

Sood et al. [11], following a similar line, detect profanity in text by employing Amazon's Mechanical Turk to label a set of comments from a social news site. The labeled dataset (represented as a *bag-of-words* where order is not important) is used to extract features such as bigrams and stems, which are used, in turn, to train an SVM. The rationale behind employing a supervised approach consists of overcoming the limitations of lexicon-based approaches, since these can fail to detect foul language by missing variations and invented or mispelled words. The authors do experiment, though, with the Levensthein distance to leverage the accuracy of the `noswearing.com` lexicon.

We attempt to go beyond binary classification by treating the aggressive text detection task as a *regression problem* and by using lexicons that, to the best of our knowledge, have not been tested for this particular task.

4 Aggressive text detection

The problem we tackle consists of automatically mapping a document d_i to an aggressiveness score sc_i. The first step towards attempting to solve this problem is defining a bounded range for the score to fall in. We consider the range $[0, 10]$ to be appropriate for this context, since it is neither extremely coarse nor extremely granular; for this range, 0 indicates no aggression and 10 indicates a strong aggression.

The second step (and probably the most difficult) concerns finding a suitable technique to produce the scoring. Because an aggressive text could be seen as intrinsically negative, we conceive aggressiveness scoring as a sub-task of *sentiment analysis*, specifically of *polarity detection*. Furthermore, given that polarity detection is mostly either lexicon-based or supervised, exploring these kinds of approaches seems reasonable; in addition, since our specific problem regards *regression*, it also seems valid to explore statistical approaches such as linear regression. Let us describe each of these candidates.

4.1 Lexicon-based approaches

Our lexicon-based approaches are, to some extent, similar to term counting using a bag-of-words model— i.e. word ordering in the text is unimportant; however, we mainly focus on detecting negative terms (or the absence of positive ones). Let us briefly provide a background on each particular approach and then describe how the aggressiveness score is generated for that approach.

4.1.1 Swear words.

Our first lexicon (which we shall refer to as "NS") is extracted from the `noswearing.com` site, which comprises a collection of offensive words and their meanings, as well as a list of variants for these words; by being open to submissions from anyone, the site resembles a wisdom-of-crowds resource, thus offering a vocabulary that appears to be well suited for our purposes. To derive a score using this lexicon, we obtain the relative frequency of offensive words for the document and normalize this frequency using the maximum that has been found in the document collection. The relative frequency f_i of offensive words for a document d_i is calculated as the proportion of swear words in d_i, such that

$$f_i = \frac{o_i}{n_i}, \tag{1}$$

where o_i and n_i are, respectively, the total of offensive words and the total of words (both for d_i). The score sc_i is finally calculated by normalizing this relative frequency with the maximum $f_{\max}$ found and multiplying the result by ten, as the normalized frequency yields a value within $[0, 1]$:

$$sc_i = (10)\left(\frac{f_i}{f_{\max}}\right). \tag{2}$$

So, for example, assume that w_1 and w_2 are swear words in a document $d_i = \{w_1, w_2, w_3, w_4\}$. In this case, $f_i = \frac{2}{4} = 0.5$; if $f_{\max} = 0.6$, then $sc_i = \frac{0.5}{0.6} = 0.83$.

4.1.2 ANEW.

This lexicon, which stands for "Affective Norms for English Words" [2], is an affective resource that has been used for measuring happiness [6]. ANEW comprises a set of 1,034 words manually scored according to three aspects or *semantic differentials* (i.e. scales whose extremes are two opposite adjectives):

1. Pyschological valence (bad-good)
2. Motivation (passive-active)
3. Domain (weak-strong)

The overall affective value for a given word (as well as the value for each of the differentials) lies within the range $[1, 9]$, where 9 is the closest to happiness. To calculate the score using ANEW, we take the overall values for the document words found in the lexicon and average them; for example, given document $d_i = \{w_1, w_2, w_3, w_4\}$, if w_2 and w_4 were found and their respective overall values turned out to be 5.0 and 8.5, the average overall value would be $\frac{(5.0+8.5)}{2} = 6.75$. Since this value reflects a degree of happiness that increases with larger values (on the contrary of our scale, where greater values are more negative) and ANEW's range slightly differs from ours, we translate the resulting averages using

$$\begin{aligned} sc_i &= \frac{(b-a)\left[(d-v_i)-c\right]}{d-c} + a \\ &= \frac{(10)\left[(9-v_i)-1\right]}{8} \end{aligned} \tag{3}$$

where $a = 0$, $b = 10$, $c = 1$, $d = 9$, and v_i is the average value obtained from document d_i; note that $[a, b]$ is our range of aggressiveness and $[c, d]$ is ANEW's range of happiness. For the example provided above, the average value $v_i = 6.75$ would be translated into an aggressiveness score $sc_i = 1.56$.

4.1.3 SentiWordNet.

The third lexicon used is SentiWordNet, which is a WordNet-based[1] tool for opinion mining. As the original WordNet, SentiWordNet contains English nouns, verbs, adjectives, and adverbs that are grouped into "synsets", i.e. sets cognitive synonyms, each expressing a distinct concept. The synsets are, as well, interlinked by means of conceptual semantic and lexical relations. SentiWordNet assigns each synset three classifications with respect to confidence, negativity, positivity, and objectivity [1]. Each synset is, therefore, associated to three numerical values: `Pos(s)`, `Neg(s)`, and `Obj(s)`. These values, respectively, indicate positive, negative or objective (neutral) polarities and fall within

[1] WordNet is available at `http://wordnet.princeton.edu`.

the range $[0, 1]$; let us note that the sum of the three associated values is necessarily 1.0, which means that each synset has a value other than zero in at least one category [1].

To calculate the aggressiveness score with SentiWordNet, we averaged the negative polarities of the document words found in the lexicon—similar to the approach followed with ANEW. However, in contrast with ANEW, the resulting average is already negative and could easily be converted to our scale by multiplying the average by ten.

An important issue to consider, though, with SentiWordnet is the presence of ambiguity; in our case, the type of ambiguity that could potentially affect our scoring is *polisemy*, i.e. words with multiple meanings. Our first attempt to handle this kind of ambiguity consists, simply, of discarding those words that have multiple negative polarities; searching for a finer disambiguation process is left for future work.

4.2 Other approaches

We also explore fuzzy, statistical, and supervised approaches. This second handful of approaches aims to combine different features or variables and is, to some extent, leveraged by the previous lexicon-based methods.

4.2.1 Fuzzy systems.

A *fuzzy system* is an expert system that works with imprecise, vague knowledge and is based, as the name suggests, on *fuzzy logic* [13]. This kind of system maps a set of given inputs to an output by means of an *inference engine* that uses a fuzzy *rule base*. To perform inferences with this rule base, the inputs are *fuzzified* and the fuzzy result is *defuzzified*; the latter process yields a "crisp" output. Fuzzy rules have an "if-then" structure that contains *linguistic variables*. A *linguistic variable* is a variable associated with a numeric variable x and whose values are *fuzzy sets*. A fuzzy set, in turn, is a set whose elements have a membership value within the range $[0, 1]$—as opposed to crisp sets where elements are either present or absent. Membership values are given by *membership functions*.

Using the brief previous framework, let us describe the design of our fuzzy system for aggressiveness scoring. While several designs have been explored, the one presented here—as we will see later—has achieved, so far, the best results for the fuzzy approach. This design considers two inputs: the document's length (total words) and the number of swear words. The output is an aggressiveness value between 0 and 1. The system, therefore, contains three linguistic variables, each defined to have five possible values or fuzzy sets (see Table 1). All fuzzy sets are represented with triangular membership functions whose parameters (start, peak, end) are determined according to the mean and standard deviation of the particular dataset being used; while the former implies redefining the function parameters for each dataset (which seems reasonable when dealing with different document collections), we believe that this criterion is better than arbitrarily defining the functions. With respect to the system's rule base, it consists of 25 rules extracted from our prior experience on the subject (see Table 2 for some examples). Fuzzification and defuzzification are, respectively, carried out with the singleton and centroid methods.

Table 1: Fuzzy sets

Document length	Number of swear words	Aggressiveness
Too short	None	Very positive
Short	Very few	Positive
Moderate	Few	Tends to be aggressive
Long	Many	Aggressive
Very long	Too many	Very aggressive

Table 2: Inference rules examples

IF document is too short and has a few swear words THEN document is positive.
IF document is short and has too many swear words THEN document is very aggressive.
IF document is moderated and has none swear words THEN document is very positive.
IF document is long and has too many swear words THEN document is very aggressive.

Since the value returned by the fuzzy system lies within the range [0,1], we only multiply this value by ten to place it in the range of our aggressiveness score.

4.2.2 Supervised learning and linear regression.

Supervised learning approaches, in contrast to fuzzy systems, act as black boxes. While this hinders their capability of explaining why a certain output was obtained, it is also true that we do not have to build an expertise-demanding knowledge base. A supervised approach, however, requires a set of *labeled examples*. Each example consists of an input (represented by a number of *features*) and its corresponding output (label). A determined amount of examples is used for *training* (learning the function that maps an input to an output) and another amount is used for *testing* (validating that the function generalizes well by using unseen examples). Neural networks are a strong representative for supervised learning; such networks aim to mimic the human brain by considering a set of *neurons* (usually spread into *layers*) connected by synaptic *weighted edges*. Neural networks learn by repeatedly adjusting these weights.

Linear regression—a classical statistics-oriented technique–also learns a model that predicts outputs based on inputs with multiple features; the model is a *linear* function (hence the name) that best fits the data.

For predicting the aggressiveness score via supervised learning or linear regression, the following set of features is considered:

1. Document length (number of words)
2. Number of offensive words (using the `noswearing` lexicon)
3. Frequency of the word "you"
4. NS score
5. ANEW score (using the 1-9 original scale)
6. SentiWordNet score

5 Experiments and results

The aim of our experiments is two-fold: on one hand, we wish to compare our candidate approaches, and on the other hand, we also wish to have a notion of which features better support aggressive text detection. For this matter, we test each approach with a set of comments extracted from *Twitter*; not only is this social network/microblogging service important and popular, but also (according to our point of view) prone to cyberbullying and harassment.

5.1 Setup

Our Twitter repository was gathered by crawling comments containing words such as "school", since this is the typical environment for cyberbullying and conversations leading to it; the collected comments belong to the English language, since the lexicons we use were made for this language (working with Spanish, which is our native tongue, is left for now as future work). From the obtained repository, we solely selected those comments that were directed towards one or more users, assuming that personal references potentially build cyberbullying attempts as well; in Twitter, directed comments are called *mentions* and are depicted with `@username`, where `username` represents the recipient of the comment. With this filtered repository, we furtherly generated two datasets: one with comments containing the word "f*ck"[2] and another containing the word "b*tch". To avoid using these words over and over again, let us refer to these datasets as, respectively, the *f-dataset* and the *b-dataset*. The reason for choosing comments with swear words obeys the intuition of finding aggression in these types of comments, while also acknowledging that both words have a certain degree of ambiguity. A summary of the repository is given in Table 3, and Table 4 presents some comment examples.

Both datasets were manually scored by four evaluators, who were instructed to place a number between zero and ten according to the perceived aggressiveness in each particular comment (zero being not aggressive and ten being very aggressive). To verify that these human judgments were similar—and, therefore,

[2]For respect, we do not show the complete swear word. The reader might guess the word we are referring to.

Table 3: Twitter repository used for this research

Classification	Number of comments
Dataset	111,381
Directed comments	12,705
f-dataset	281
b-dataset	110

useful to our purposes—, we carried out ANOVA (see Table 5). Comments were discarded until the test was past, leaving the f-dataset with 174 comments and the b-dataset with 69 comments; let us note that the comments removed were the ones where there was no uniform judgment (e.g. one evaluator placed a mild aggression score and another placed a strong aggression score). This reveals the degree of complexity for the task, since not even humans agree in a percentage of the cases.

Table 5: Results for analysis of variance

ANOVA	f-dataset	b-dataset
Population (n)	174	69
Evaluators (a)	4	4
Independent evaluator subtotal [A]	4610.2	1946.1
Sum of [A] [T]	4275	1685.2
Individual value [Y]	14271	3868
Square sum between groups	334.9	260.9
Square sum within groups	9660.8	1921.9
Degree of freedom between groups	3	3
Degree of freedom within groups	692	272
Mean square between groups	111.6	86.9
Quadratic mean in group	13.9	7.1
F Ratio (F)	7.9	12.3
Comparison between observed F vs. Distributed F	2.6 / 3.8	2.7 / 3.8

The final comments on both datasets were pre-processed by eliminating punctuation marks, changing every word to lower case, and utilizing *regular expressions* to: correct misspelled words (for instance, “biatch” or “biotchhh”), expand acronyms such as “OMFG”, and separate words with swearing (e.g. the username @muppyb*tch would be broken down into @muppy and b*tch). We

Table 4: Examples of comments (swear words are partly censored).

Comments
@F*ckCrystal i gotta go to school at 5 so idk if you wanna chill after that??
@ParishRory @Sh4niqua pffffft hah no way shes a f*cking bully. I'm actually scared of her

also translated emoticons such as `:)`, `:(`, and `:@` to affective terms like "happy", "sad", or "angry".

Before showing and discussing results, let us note that we selected the *multilayer perceptron* neural network as our supervised learning approach; the parameters for this neural network (as well as for linear regression) were the default used by the WEKA toolkit. Training and testing were performed using a cross-validation of ten folds. In addition, different runs were performed using all attributes, all atributes minus no. 3 (you's), only attributes 1,2, and 4 (length, badwords, and NS score), and only attributes 1 and 2. Let us respectively refer to these variants as *6-attribute*, *5-attribute*, *3-attribute*, and *2-attribute*.

5.2 Results and discussion

Each approach was evaluated using the two datasets. For comparison, we employed the Mean Squared Error (MSE), which is calculated as $(x-y)^2$, where x is the average human score and y is the score obtained using a particular approach. To have a clearer view of results, we also introduced a baseline method, which consisted of randomly-generated scores; such random scores were generated 30 times and then averaged.

Our results are shown in Figures 1 and 2; Table 6, more precisely, depicts all errors. The overall best approach (lowest MSE) was the 2-attribute linear regression, followed by the 3-attribute neural network, the fuzzy system, the NS lexicon, SentiWordNet, and finally ANEW. Interestingly, this last approach was worse than the baseline; we believe this may be due to the presence of slang and informal text, as well as to some kind of ambiguity. If we contrast single-source approaches (lexicons) vs. multiple-source approaches (supervised, statistical, and fuzzy), there is also an important difference; in that sense, those methods combining different variables or features seem to work better than methods with a single type of information. Furthermore, if we contrast the results obtained per dataset, we may note that the f-dataset in general obtained smaller errors than the b-dataset; for statistical and supervised approaches, this could be due to the size of the dataset (more examples available to train). Within the lexicon-based methods, not surprisingly, the best results were obtained by the NS approach; while this could be partly due to the datasets (chosen by searching for swear words), we believe that the strength of the approach rather lies on the close relationship that exists between aggressiveness and profane language. We could also question whether aggressiveness could be conceived without taking into account this kind of language. In that sense, the presence and number of swear words in the text could act as a key feature.

Table 6: MSE Results

Approach	f-dataset	b-dataset	Average
NS	5.2	7.2	6.23
ANEW	16.1	33.9	24.95
SentiWordNet	11.2	8.4	9.8
Fuzzy system	4.8	6.1	5.5
Neural network	4.2	6.2	5.2
Linear regression	**3.6**	**5.9**	**4.8**
Baseline	15.6	20.1	17.9

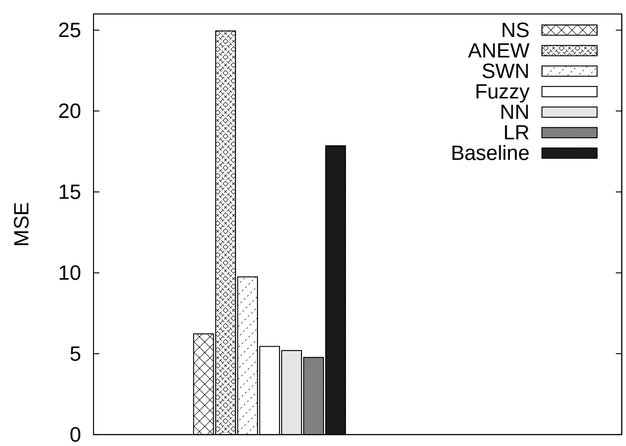
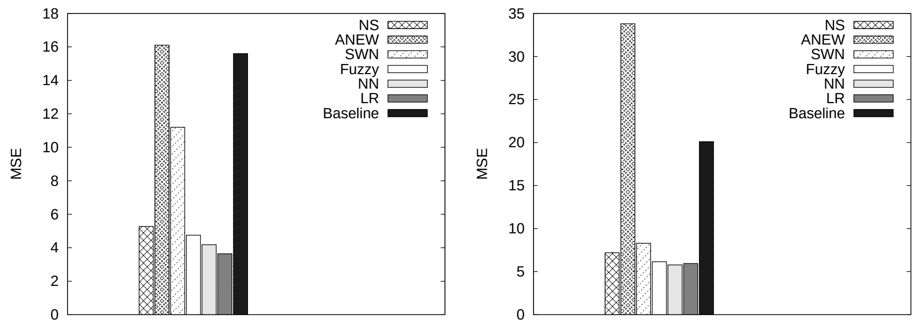


Figure 1: Average mean squared error (MSE). NS= `noswearing.com` lexicon, SWN= SentiWordNet, NN= Neural Network, LR= Linear Regression.

By drilling down the obtained results, it is also important to note that the hardest cases for all approaches were the ones with a high degree of aggressiveness; we believe this is due to several reasons. On one hand, these comments are more scarce than the rest, which implies less examples to train or characterize. On the other hand, there could be aspects in those comments that need to be considered, such as the underlying emotions, intentions, and context. It would also be interesting to include weight assignment for swear words and evaluate if this change impacts some of the results.



(a) MSE for the f-dataset (b) MSE for the b-dataset

Figure 2: Mean squared error by dataset.

6 Conclusions and future work

In the present work, we have tackled aggressive text detection as a regression problem that consists of mapping a document to an *aggressiveness score*; to the best of our knowledge, state-of-the-art methods tend to cast this issue as a binary classification problem. We have defined a simple scale that ranges from zero to ten (where ten is the most aggressive) and assumed, as well, that aggressive text detection is a sub-task of sentiment analysis that is closely related to document polarity detection. Taking the former into account, we proposed and explored lexicon-based, supervised, fuzzy, and statistical approaches, which were tested over a Twitter repository. Our results show that linear regression seems to be a solid candidate for scoring the documents, and that the use of profane language (swear words) seems also to be a key feature for the task.

Future work includes refining our approaches to better handle difficult cases (e.g. creating other designs for the fuzzy system), testing more supervised approaches (SVM, for instance), using a larger dataset, working with documents in Spanish, and building a framework for cyberbullying automatic identification.

References

[1] Análise de sentimentos no Twitter utilizando SentiWordNet, 2011. Proposta de Trabalho de Graduação.

[2] Margaret M Bradley and Peter J Lang. Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical report, Center for Research in Psychophysiology, University of Florida, 1999.

[3] Marilyn A Campbell. Cyber bullying: An old problem in a new guise? *Australian journal of Guidance and Counselling*, 15(1):68–76, 2005.

[4] M. Dadvar, F.M.G. de Jong, R.J.F. Ordelman, and R.B. Trieschnigg. Improved cyberbullying detection using gender information. In *Proceedings*

of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012), pages 23–26, Ghent, 2012. University of Ghent.

[5] Karthik Dinakar, Roi Reichart, and Henry Lieberman. Modeling the detection of textual cyberbullying. In *The Social Mobile Web*, 2011.

[6] Peter Sheridan Dodds and Christopher M Danforth. Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies*, 11(4):441–456, 2010.

[7] Dorothy L Espelage and Susan M Swearer. Research on school bullying and victimization: What have we learned and where do we go from here? *School Psychology Review*, 2003.

[8] Alistair Kennedy and Diana Inkpen. Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, 22(2):110–125, 2006.

[9] Bing Liu. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167, 2012.

[10] Vinita Nahar, Li Xue, and Pang Chaoyi. An effective approach for cyberbullying detection. *Communications in Information Science and Management Engineering*, 2012.

[11] S Sood, Judd Antin, and Elizabeth Churchill. Using crowdsourcing to improve profanity detection. In *AAAI Spring Symposium Series*, pages 69–74, 2012.

[12] Fabio Sticca and Sonja Perren. Is cyberbullying worse than traditional bullying? Examining the differential roles of medium, publicity, and anonymity for the perceived severity of bullying. *Journal of youth and adolescence*, pages 1–12, 2012.

[13] L.A. Zadeh. Fuzzy sets. *Information and control*, 8(3):338–353, 1965.