"How Bullying is this Message?": A Psychometric Thermometer for Bullying

Parma Nand, Rivindu Perera, Abhijeet Kasture
School of Engineering, Computer and Mathematical Sciences
Auckland University of Technology
Auckland, New Zealand
{pnand,rperera,akasture}@aut.ac.nz

Abstract

Cyberbullying statistics are shocking, the number of affected young people is increasing dramatically with the affordability of mobile technology devices combined with a growing number of social networks. This paper proposes a framework to analyse Tweets with the goal to identify cyberharassment in social networks as an important step to protect people from cyberbullying. The proposed framework incorporates latent or hidden variables with supervised learning to determine potential bullying cases resembling short blogging type texts such as Tweets. It uses the LIWC2007 - tool that translates Tweet messages into 67 numeric values, representing 67 word categories. The output vectors are then used as features for four different classifiers implemented in Weka. Tests on all four classifiers delivered encouraging predictive capability of Tweet messages. Overall it was found that the use of numeric psychometric values outperformed the same algorithms using both filtered and unfiltered words as features. The best performing algorithms was Random Forest with an F1-value of 0.947 using psychometric features compared to a value of 0.847 for the same algorithm using words as features.

1 Introduction

Harassment and bullying as common forms of unacceptable behaviour have been topics in psychological research for a long time. Harassment aims to discriminate people on the basis of race, sex, disability etc. and bullying aims to intimidate people. The use of the terms "cyberharassment" and "cyberbullying" is relatively new. Both terms involve the application of modern technologies to support negative human behaviour. The ubiquitousness and affordability of modern technology makes it easy for organisations and individuals to easily communicate via e-mails, chat rooms, message boards and social media which generates a huge amount of public and private data containing information reflecting the pulse of the society. Unfortunately, this accessibility of the technology has also created a forum for the expression of negative human emotions, some of which also gives rise to tragic outcomes such as suicides, self harm and mental illnesses. The aim of our research is to create a framework for detecting cyberharassment from textual communication data, so that systems can be implemented to catch and eliminate harassing texts before it is able to inflict harm.

Bullying is a relationship issue that could result in significant emotional and psychological suffering with some even resulting in suicides (Boyd, 2007). Cyberbullying is even worse because it can follow the victim everywhere, happen at anytime and it is frequently anonymous. The number of victims in all age groups is growing worldwide. It is urgent to progress the research in this area in order to develop advanced methods for fast detection and prevention of cyberbullying cases. This involves analysis of a huge amount of social network textual data which is largely unstructured and chaotic.

A range of challenging tasks are associated with cyberharassment research. For example, bullying is to a great extent connected to the nature of the victim rather than solely on the language and topic of text. It generally involves the use of known weaknesses of victims that they cannot change. Texts can still contain profanities and include sensitive topics without being bullying or any psychological effect on the

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 695–706, Osaka, Japan, December 11-17 2016.

victim. In order determine whether if cyberbullying occurred, we would need to attain information on the victim's reaction and classify this reaction. This is a challenging task in itself and not part of our work. The context or background knowledge can also determine if a text is bullying or not. It is possible for individuals in close friendship to communicate using profanities and on sensitive topics, without causing a negative psychological effect. The actual context is normally not included in texts since it is difficult to determine, especially for short pieces which is typical of social media type communications.

To be able to computationally detect cyberbullying, one would need texts on time-line based conversational threads between individuals, which is difficult to obtain due to privacy and propriety implications. Although there has been an abundance of studies on cyberbullying from the social and psychological perspectives, good computational systems to identify cyberbullying are rare. The existing computational studies (Dinakar et al., 2011; Yin et al., 2009; Xu et al., 2012b; Van Royen et al., 2015; Cortis and Handschuh, 2015; Squicciarini et al., 2015; Han et al., 2015; Kansara and Shekokar, 2015) use predominantly keywords as features in classification algorithms. Those approaches deliver results which are good indicators of harassment and they provide text which can be processed further to find bullying cases. Identification of harassment in singleton blogging texts would be extremely useful on social media platforms as a monitoring tool. Once a blog has been found to be harassing, the corresponding conversations could then be monitored more closely over a certain time in order to identify the progress from cyberharassment to cyberbullying.

In this paper, we apply a well established theory from psychology integrated in a freely available software tool to develop a framework with the goal to classify pieces of texts as bullying. Instead of using a lexicon of profanities, the framework applies a much richer dictionary of words and weights, referred to as psychometric measurements in the psychology literature, to create a rich set of numeric features which are then used four different classification algorithms. We train four different classifiers using two types of filtered and unfiltered inputs (high correlation features and annotated text). All classifiers delivered reasonable results, however, the "Random Forest" classifier achieved the highest precision rate.

1.1 Psychometric Analysis

Harassment and bullying are manifestations of negative emotions which are described as *variables* in psychology (Browne, 2000). In many scientific experiments, we obtain exact measurements of well defined features and then we normally derive answers for correlated questions under investigation. It is not straightforward to find answers to questions that deal with aspects of human psychology such as bullying and harassment. Features related to psychological constructs are typically not clearly defined and cannot be directly measured. Additionally, indirect determinations of values are often subject to substantial measurement errors and varying degrees of correlation with the question under investigation. Psychologists declare those features as latent or hidden variables. *Latent variables* are those that are indirectly measured by deducing relations between manifest, or *observed variables* (Browne, 2000).

Psychometrics is a discipline that deals with the quantification and analysis of capacities and processes in human brains, which often manifest as latent variables. Psychometrics deals with the construction of procedures for measuring psychological constructs and the development of mathematical or statistical models for the analysis of the psychological data. Typically, psychological constructs are multi-variate consisting of a large proportion of latent variables. Researchers generally measure observed variables and then find correlations with the latent variables under investigation.

Psychometrics has been used in applications where measurements of some aspects of human psychology are required. For instance, it has been used to determine an individual's personality alignment to a given set of characteristics. In this study, the psychometric variables are measured to evaluate strengths and weaknesses of a person for specific tasks. The results have been mapped to their cognitive abilities and general social behavioural style as described by Kline (2013). Many companies perform psychometric evaluations on candidates to identify potential matches for specific job roles. It has also been used to identify individuals' suitability towards specific career paths using psychometric techniques on observed variables.

An individual's writing style is an example of an observed variable, hence its psychometric evaluation

would be able to provide insights into the individual's latent characteristics on a range of psychological processes. Further, psychometric evaluation performed on texts collectively generated by various people on a similar topic can provide insights into the psychological processes of this group as a whole. Twitter, for instance, provides the ability to use "hashtags" which allows multiple users to easily write on a designated topic. This makes it easy to collate Tweets on a topic for the purpose of carrying out psychometric evaluations such as opinion mining on the topic. For example, Poria et al. (2014) provide dependency based rules for sentiment analysis. Apart from binary classification of opinions, more advanced psychometric evaluations can even distinguish between different degrees of individual opinions on a topic.

1.2 Related Work

Detecting cyberharassment is a special case of text classification which is an older research area compared to research on cyberharassment. Text classification involves use of generic machine learning algorithms to classify texts into two or more categories using features extracted from the words in the text. In order to be able to extract features there maybe various types of preprocessing tasks such as Part-of-Speech tagging, stop word removal, named entity recognition, etc., applied which differentiates the various systems and adapts them better for the type of classification application at hand. Text classification techniques developed in the areas of sentiment analysis and opinion mining are especially applicable to cyberharassment detection as they also deal with human psyche expressed as online texts. For example, Bollen et al. (2011) report the results of a text classification framework applied to support business decision. This paper presents the results for analysing 10 million Tweets from three million users over a 10 month period to predict changes in the stock market. The authors aim to answer whether the analysed mood expressed in the Tweets from a cross section of three million people is related to the Down Jones Industrial Average(DJIA). The authors applied the following two mood tracking tools:

- OpinionFinder, a system that processes Tweets and separates them into positive versus negative sentiments
- Google-Profile of Mood States (GPOMS), a system that classifies Tweets based on the mood categories; happy, vital, alert, calm, sue, kind.

Granger Causality Analysis and fuzzy neural network classifiers were used for cross-validation of the resulting mood time series against large public events such as US presidential election and Thanksgiving in this paper. The paper reports a correlation of the changes in DJIA value with predicted values from the classifier with an accuracy of 87%.

Cortis and Handschuh (2015) report cyberbullying analysis in the backdrop of Tweets related to two top trending world events in 2014, namely Ebola virus outbreak in South Africa and shooting of Michael Brown in Ferguson, Missouri. These events generated bullying comments on Africa people with no connection to Ebola virus and racist comments regarding the shooting for each of the events respectively. The authors of this paper used only limited number of key terms (*whore, hoe, bitch, gay, fuck, ugly, fake, slut, youre, thanks*) to classify Tweets against ground truth classifications as annotated by trained curators. The combined precision values for key terms such as "Whore" ranged from 0.75 to extremely low values of 0.2 for key terms such as "youre" with an average precision of approximately 0.53.

Dadvar and de Jong (2012) also present a cyber bullying detection framework based on MySpace posts. This paper uses a large dataset of 2,200 posts annotated by students and use a SVM classifier. The interesting aspect of this work is that they used four types of features. All the profane words were treated as in one category and the ratio of foul words to the length of the post was used as the feature for the classifier.The other features used were second person pronouns, all other pronouns and TF-IDF all the words in each post. Additionally, the posts were also split by gender based on the hypothesis that bullying characteristics are between the genders. The reported precision value for male specific corpus was 0.44 and for female specific corpus was 0.40. However, the respective recall values were 0.21 and 0.05, which is rather low even after considering the data set size of 2,200 posts. The performance numbers from this

study in addition to similar numbers from others such as (Ptaszynski et al., 2010; Nandhini and Sheeba, 2015; Xu et al., 2012a) was the motivation for us to use the whole text instead of a keyword list based approach to detect cyberharassment in the previous works. The framework presented in this paper uses the psychometric analysis of all the work in the text and converts them to numeric values before using them as features in the classification algorithms. The next section describes this framework in detail.

2 Proposed Framework

In this paper, we describe a framework for processing Tweets to generate a model for the prediction of cyberbullying. Four main groups of experiments were done in order to determine the best performing classifier for the different forms of input data. The following variations of the Tweet data were tested after cleaning it for extra punctuations and repeated words. Various combinations of

1. All words.
2. Words with stopwords removed.
3. Words above a threshold correlation value with 2 types of filters, Correlation Feature Selection (CFS) and Infogain Attribute Evaluator (IAE).
4. Psychometric values for all words.
5. Psychometric values for stopwords removed.
6. Psychometric values for filtered words.

2.1 Architecture

Figure 1 shows a schematic representation of the framework. We divided our model into the following phases:

1. Extraction of Tweets with typical keywords for bullying using the TAGS tool.
2. Pre-processing of archived Twitter data.
3. Human annotation of Tweets to establish ground truth.
4. The LIWC2007 software generates a vector for each Tweet and the results are saved in two EXCEL files, corresponding to true positives and true negatives.
5. Vectors are filtered (using two different techniques) to reduce the number of attributes before classification.
6. Combined files (filtered or unfiltered) are used to train four machine learning classifiers in WEKA.
7. Evaluation of the classification results.

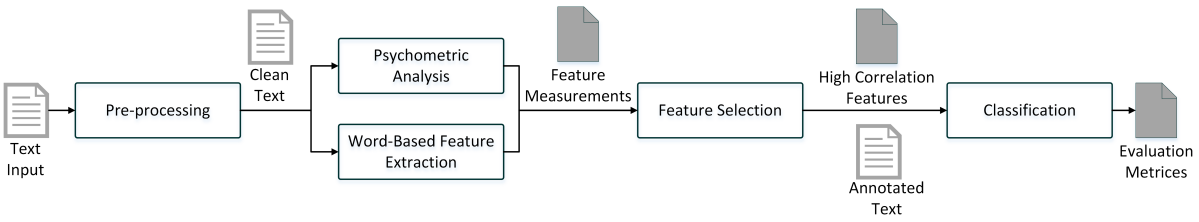

Figure 1: Schematic representation of the Architecture

2.2 Extraction of Tweets

We used the Twitter Archiving Google Spreadsheet[1] (TAGS) available for the public to explore opportunities for the exploitation of the huge amount of data generated everyday by the Twitter community.

2.3 Data preparation/Pre-processing

Tweets have a size restriction of 140 characters and typically originate from mobile devices, and due to small keypads they contain excessive amounts of noise. Apart from the "typo" errors, a range of other types of noise are usually contained in Tweet messages, such as, extra space and non-ASCII characters, due to the nature of the platform. In order to be able to extract the maximum possible semantics, we eliminated the following types of noise using techniques described by Nand et al. (2014).

Word Variations - e.g., tmro and 2moro replaced by tomorrow

Multi Word Abbreviation - e.g., lol replaced by laugh out loud

Slangs - e.g., gonna replaced by going to.

One of the major hurdles in researching online harassment is access to data, which, by the nature of the purpose, is usually private to an individual or to a closed group. Twitter, however, does have publicly available conversational data exhibiting harassment characteristics. For this research we downloaded a total of 2500 public Tweets using the TAGS archiving tool to generate Google Spreadsheets over a period of 2 months. The Tweets were retrieved by entering typical keywords for bullying as recommended in psychology literature (Ortony et al., 1987; Cortis and Handschuh, 2015; Squicciarini et al., 2015; Browne, 2000; Ybarra, 2010).

Keywords used to download Tweets: nerd, gay, loser, freak, emo, whale, pig, fat, wannabe, poser, whore, should, die, slept, caught, suck, slut, live, afraid, fight, pussy, cunt, kill, dick, bitch.

After removing the duplicates and ReTweets we had a sample of 1689 instances containing one or more of the keywords. In addition, @Usernames. #Hashtags and Hyperlinks were removed because the LIWC (Linguistic Inquiry and Word Count) - tool accepts only plain ASCII text files. The cleaned texts were then manually classified by a set of 3 trained annotators into either "cyberbullying" and "non-cyberbullying". The sample was divided into three lots containing 563 instances each. Each lot was assigned to two different annotators which meant that each Tweet was annotated twice by two different annotators. The annotators tagged the instances based on following guide lines:

Character Attacks: the reputation, integrity and morals of an individual are targeted with the purpose of defamation.

Competence attacks: bullies denigrate individuals ability to do something.

Malediction: an attack in which bullies curse and express a wish for some type of misfortune or pain to materialize in an individuals life.

Physical appearance: targeted on an individuals look and bodily structures. Typically, physical attributes of humans are found to shape and develop their personality and social behavioral relations. Due to the need of an individual to be socially accepted, these types of attacks make victims feel socially neglected making a long lasting negative impact on their self-esteem.

Insults: profanity is used as an attack wherein bullies use extremely offensive language that typically include foul, lewd, vulgar language in addition to swearing and cursing words.

Verbal abuse: includes false accusations or blames, extreme criticisms and judgements about an individual and/or statements that negatively define the victim.

[1] https://tags.hawksey.info/

Teasing: hurtful in nature and done as a spectacle for others to witness resulting in harassment and humiliation of the victim. It is perceived as a form of emotional abuse.

Threats: generally anonymous in cyberbullying. Due to this anonymity victims tend to live in constant fear that leads to long-lasting depression, low self-esteem and delinquent behaviours.

Name-calling: bullies use denigrating, abusive names and associate them to the victims leaving them humiliated in front of others.

Mockery: pass comments on victims making them feel worthless, disrespecting them and make fun of them in front of others. Escalated form of mockery leads to low self-esteem of the victims.

In order to ensure that all true positives were identified each of the three sets of Tweets were annotated twice by two different annotators with a Cohen's kappa of 0.833. This resulted in a total of 427 instances as harassing Tweets (true positives). In order to further validate the true positives, we did a second round of annotation by randomly dividing the 427 true positives into three lots, and had two annotators to again check each true positive instance. If a true positive instance was rejected by two annotators in the second round we rejected that instance and assigned it as non-cyberbullying or true negative. If a tweet was annotated as bullying by one annotator and non-bullying by the second, then it was looked at together and an agreement was reached. After the second round we had a total of 376 true positive instances which was taken as the ground truth.

2.4 LIWC2007, Psychometric Analysis

Linguistic Enquiry Word Count (LIWC) is a free text analysis application, originally designed to study emotions in text files. It started with counting words in psychology which were relevant to certain categories and it has been developed into a continuously improving software tool. At the beginning, human emotion words were categorized only into negative and positive words. The system architecture of the LIWC too is described in (Pennebaker et al., 2015) and the psychological categories can be found in (Tausczik and Pennebaker, 2010). The default dictionary of LIWC2007 is composed of 4,500 "dictionary words" grouped into four high level categories. The input text is a set of so-called "target words". LIWC2007 processes text files in one of many provided formats and writes the analysis in one output file. If a target word has been found in the dictionary, it will be assigned to one or more categories and the count of each of the categories to which the word belongs is incremented. For example, the word "cried" is part of five categories: sadness, negative emotion, overall affect, verb and past tense verb. The target words are the elements of the pre-processed Tweets in our study, where each Tweet is a unique instance. The output is a vector with 67 numeric values per Tweet corresponding to the number of counts for each of the word category divided by the total number of assignments. The average number of words per Tweet in all of the used, 1313 Tweet, was 18 and 89% of our target words have were found to be in the dictionary. The average number of words in a cyberbullying Tweet was 13 out of which an average of 11 words were assigned to the psychometric word categories. Mostly internet slang words or names of locations were not found in the dictionary.

3 Results and Evaluation

The graph in Fig. 2 shows the average psychometric measurement values for a subset of all 67 features generated by the LIWC tool for the Tweets manually identified as cyberbullying (blue or left line) and those that are not identified as cyberbullying (red or right line). The differences between the calculated values (length of the red and the blue line) for some word categories such as "you", "swear", "negemo", "anger", "bio" and "death" are relatively large and we can assume that the discriminatory power for true positives would be high for these features. But the results have shown that the classifier accuracies dropped by about 6% when the filters reduced the number of features to these features with high correlation. Another observation is that the value for the "sexual" category is almost equal for positive and negative instances because the downloaded Tweets have been selected based on a majority of words

belonging with the "sexual" category. From this we can infer that the words which are elements of the "sexual" category is not necessarily a good predictor of cyberbullying.

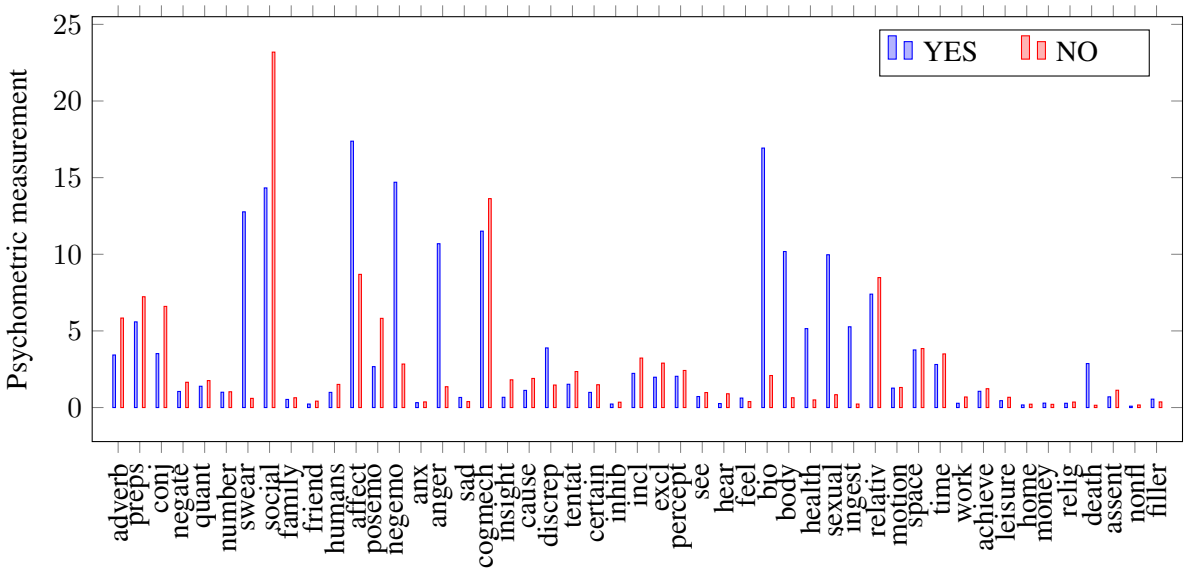

Figure 2: Graph showing the average psychometric measurements for positive and negative instances

The four classifiers have been tested in three lots. The first experiment used all 67 numerical values per instance generated by the LIWC software. Table 1 shows the numerical results for precision and recall for our selected classifiers. The second column in Table 1 provides the results of the cross validation of our trained model using the training data set and the third column shows the results for the test data. We used the usual split of available data, that means 66% for the randomly selected training set and the remaining 34% for the test set. The precision values for the Random Forest classifier are almost equal. SMO achieved the largest value for precision of the training set compared to the other classifiers, but the Random Forest classifier delivered a higher value for the test set. The Random Forest classifier for identifying cyberbullying Tweets had the best performance in this experiment.

Classifier	Cross-Val. 10 Training	Split (66-34) Testing	Results"YES"
Random Forest	0.984	**0.983**	Precision
	0.963	**0.935**	Recall
SMO	0.986	0.965	Precision
	0.912	0.902	Recall
Multilayer Perceptron	0.963	0.873	Precision
	0.912	0.894	Recall
J 48 Decision Tree	0.947	0.935	Precision
	0.941	0.935	Recall

Table 1: WEKA Classifier Outputs for Unfiltered Inputs

The second experiment used only the filtered attributes for classification. Table 2 shows the results for the two different filters for the training set (left two columns) and the test set (right two columns). Precision and recall values for the test data are significantly lower Random Forest, Multilayer Perception and J 48 Decision Tree. The SMO classifier performs best with respect to precision and the ING filter delivered slightly better results for the this classifier. Both filters result in equal results for precision of the training data set when the the Random Forest classifier has been selected, but ING performed slightly better for the test data. Only the J48 Decision Tree classifier delivers for the test data the same results for with respect to filtered and unfiltered inputs, for all other classifiers the achieved values (shown in Table 1 and

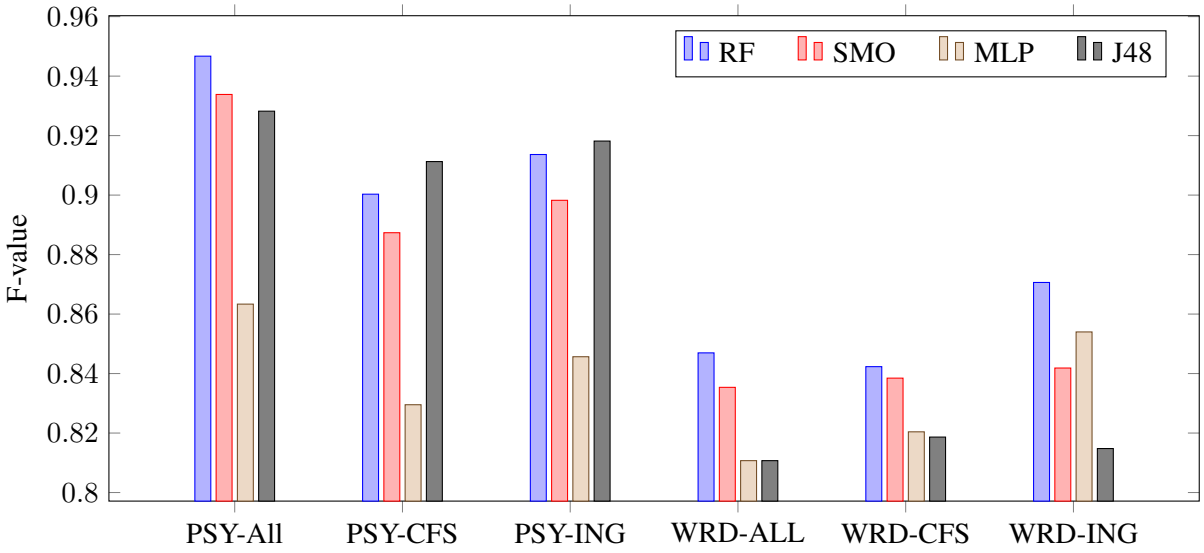

Figure 3: Graph showing the comparison of the F-values for Psychometric versus Words as Features in four classifiers.

Table 2) are equal or better for unfiltered input data.

	Cross-Val. 10		Split (66-34)		
Classifier	CFS	ING	CFS	ING	Results"YES"
Random Forest	0.978	0.978	0.967	0.951	Precision
	0.960	0.952	0.943	0.943	Recall
SMO	0.982	0.986	0.965	0.974	Precision
	0.894	0.910	0.886	0.902	Recall
Multilayer Perceptron	0.951	0.964	0.940	0.866	Precision
	0.939	0.918	0.894	0.894	Recall
J 48 Decision Tree	0.954	0.944	0.932	0.935	Precision
	0.931	0.941	0.894	0.935	Recall

Table 2: WEKA Classifier Outputs for Filtered Inputs

The third experiment was done to determine the effectiveness of the psychometric values generated by the LIWC tool, used as features for the prediction of cyberbullying. Table 3 provides the values for precision, recall and F-value for our test data, filtered and unfiltered applying psychometric features in the upper part of this table. The lower part of the table is a summery of values for the case when we use cleaned Tweets as inputs for the same filtering and classification procedures without applying the LIWC tool to compare the results. The integration of psychometric features produced for all classifiers (filtered or unfiltered attributes) better values. The outcome of our third experiment is a clear vote for transferring the clean text into the 67 LIWC attributes to identify cyberbullying Tweets.

The computed F-values in Table 3 are plotted as a graph in Fig. 3. Table 3 shows that the best performing algorithm was Random Forest with 0.98 precision and 0.91 recall using all 67 psychometric features. The filtered subsets of attributes delivered an approximately 5% drop in accuracy for the Random forest classifier and similar results for the other algorithms. Table 3 also shows that all algorithms performed better with psychometric features compared to word features. For example, the Random Forest precision is 0.983 using psychometric features compared to 0.869 using word features. The graph in Fig. 3 shows that all classifiers using psychometric features perform better than the same classifiers

Psychometric Value Features	**RF**	**SMO**	**MLP**	**J48**
All Features				
Precision	**0.983**	0.968	0.875	0.956
Recall	0.913	0.902	0.852	0.902
F-value	**0.947**	0.934	0.863	0.928
CFS				
Precision	0.926	0.912	0.859	0.926
Recall	0.876	0.864	0.802	0.897
F-value	0.900	0.887	0.830	0.911
ING				
Precision	0.932	0.925	0.863	0.936
Recall	0.896	0.873	0.829	0.901
F-value	0.914	0.898	0.846	0.918
Word Features	**RF**	**SMO**	**MLP**	**J48**
All Word Features				
Precision	0.869	0.859	0.826	0.826
Recall	0.826	0.813	0.796	0.796
F-value	0.847	0.835	0.811	0.811
CFS				
Precision	0.875	0.869	0.829	0.836
Recall	0.812	0.810	0.812	0.802
F-value	0.842	0.838	0.820	0.819
ING				
Precision	0.889	0.874	0.856	0.828
Recall	0.853	0.812	0.852	0.802
F-value	0.871	0.842	0.854	0.815

Table 3: Table showing the evaluation metrics for Psychometric and Words as features used on Random Forest, SMO, MLP and J48 classifiers

using words. There were two marked observations from the experiments; Firstly, All four of the chosen classifiers perform better with all of the features rather than a subset of higher correlation features. This was true for both the techniques used to filter the features and for both types of features, the psychometric values as well as words used as features. Secondly, the performance of the classifiers were higher for all the classifiers using psychometric values compared to use of raw words as features. Again, this was true for all the cases of using all features as well as subset of features using the two filtering algorithms.

Many text classification approaches have been performed with Support Vector Machines (SVM) (SMO is a special version of SVM) and the results have shown in many cases that SVM's are the best classifiers for text (Liu et al., 2005; Lee et al., 2012; Isa et al., 2008; Simanjuntak et al., 2010). In our case, the results show that the Random Forest classifier performed slightly better than SVM in both cases (psychometric numeric measurements as well as words used for features). Its noteworthy, to mention that the J48 algorithm also did comparatively well with an F-value over 0.9 for psychometric measurements used as features. While Random Forest, Support Vector Machine and J48's performance are quite close in most of the cases, the Multi Layer Perceptron performed is consistently lower with F-values around 0.8. The results have shown that the unsuitability of MLPs for cyberbullying, reaffirmed the suitability of SVMs for text classification. However, in addition, the results show that decision tree based algorithms, Random Forest and J48, can also perform well in short text classification, which is contrary to most of the previously reported results.

4 Discussion on Cyberharassment and Future

The issue of cyber harassment in this study as well as other studies have shown that it is complex, both in terms of definition of the problem as well as finding a solution for the problem. Almost all of the prior works (Cortis and Handschuh, 2015; Nandhini and Sheeba, 2015; Kansara and Shekokar, 2015; Xu et al., 2012a; Dinakar et al., 2011; Dadvar and De Jong, 2012; Han et al., 2015) on cyberbullying have worked on the generic problem of cyberharassment, however have referred to them as cyberbullying. However, Hosseinmardi et al. (2015) have analyzed the issue in detail and distinguished between targeted bullying, which has been referred to as cyberbullying in this paper, and general cyber aggression, referred to as cyberharassment in this paper. The broader definition of cyber aggression includes any form of digital media use to intentionally harm another person or persons. This includes targeting individuals by name calling, flaming, denigration, exclusion, etc. as well as indirect forms such as use of profanities, slangs and comments which are indirectly applicable to the individual's group, or the choice(s) made by the individual. Hosseinmardi et al. (2015), provide two additional criteria for cyberbullying as an imbalance of power between the aggressor and the victim. Their second criteria is the repetition of the act over time. The imbalance of power could be in several forms such as physical, psychological or social. For instance, an individual who is more popular bullying a less popular one or one who is physically stronger bullying a weaker one. This characteristic adds an additional level of complexity for automatic detection systems since all forms of the power imbalance is rarely known for members on social media platforms. The second criteria of repetition poses yet another level of complexity, which would require nested layers of data collection and analysis corresponding to threads of conversation rather than contents of individual posts. Research on cyberharassment will continue to be based on contents of individual posts unless the social media sites relax the data access policies and give access to thread based data. Nevertheless, research has continued to progress in detecting different forms of cyberbullying cases, especially in light of increased number of cyberbullying cases resulting in suicides and other forms of self harm. In addition to the text based platforms such as Twitter, there are several social media platforms, such as MySpace and Instagram which use a multi-modal communication model. Use of pictures and videos in addition to texts gives additional ammunition to potential bullies and presents an even greater challenge for bullying detection researchers. In spite of these challenges researchers (Han et al., 2015; Hosseinmardi et al., 2015) have attempted to work on systems to detect bullying on multi-modal social platforms by using a combination of pictures and text to detect true positives with reasonably good accuracy. For instance, Hosseinmardi et al. used a combination of text, pictures and user metadata to classify Intagrams with a recall of 76% and a precision of 62% using a MaxEnt classifier. The authors used linear SVM classifiers text based n-grams with stop word removal and used crowd sourced CrowdFlower website to establish ground truth. Although the use of the images in this study was at a basic level with categorization of images into categories such as Person, Drugs, Car, Nature and Celebrity as features for the classifier, this is a step in the right direction towards processing multimedia blogs. A lot more effort is required before it would be possible to build systems to detect harassment on multimedia social media platforms.

5 Conclusion

The results presented in this paper show that firstly it is possible to detect cyberharassment with fairly good precision. Secondly, we have shown that the use of psychometric measurements from the LIWC tool as features delivers slightly better results compared to the use of words for all of the algorithms tested. Further, out of the four classifiers tested, the Random Forest classifier proved to be the best performing both with words as classifiers as well as with psychometric measurements. The psychometric measurements uses the classification of all of the in the text which implies that cyberharassment is not only a function of profane words, but use of other category of words such as use of pronouns also determine whether a piece of text is harassing or not. As future work, we plan to use psychometric measurements with thread based social media texts instead of individual pieces of texts as this will be able to better determine the level of bullying rather than harassment as has been the case in this study.

References

Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

Danah Boyd. 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 digital media volume*, pages 119–142.

Michael W Browne. 2000. Psychometrics. *Journal of the American Statistical Association*, 95(450):661–665.

Keith Cortis and Siegfried Handschuh. 2015. Analysis of cyberbullying tweets in trending world events. In *Proceedings of the 15th International Conference on Knowledge Technologies and Data-driven Business*, page 7. ACM.

Maral Dadvar and Franciska De Jong. 2012. Cyberbullying detection: a step toward a safer internet yard. In *Proceedings of the 21st International Conference on World Wide Web*, pages 121–126. ACM.

Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of textual cyberbullying. *The Social Mobile Web*, 11:02.

Richard Han, Qin Lv, and Shivakant Mishra. 2015. Analyzing labeled cyberbullying incidents on the instagram social network. In *Social Informatics: 7th International Conference, SocInfo 2015, Beijing, China, December 9-12, 2015, Proceedings*, volume 9471, page 49. Springer.

Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishr. 2015. Prediction of cyberbullying incidents on the instagram social network. *arXiv preprint arXiv:1508.06257*.

Dino Isa, Lam H Lee, VP Kallimani, and Rajprasad Rajkumar. 2008. Text document preprocessing with the bayes formula for classification using the support vector machine. *IEEE Transactions on Knowledge and Data engineering*, 20(9):1264–1272.

Krishna B Kansara and Narendra M Shekokar. 2015. A framework for cyberbullying detection in social network. *International Journal of Current Engineering and Technology*, 5.

Paul Kline. 2013. *Handbook of psychological testing*. Routledge.

Lam Hong Lee, Chin Heng Wan, Rajprasad Rajkumar, and Dino Isa. 2012. An enhanced support vector machine classification framework by using euclidean distance function for text document categorization. *Applied Intelligence*, 37(1):80–99.

Tie-Yan Liu, Yiming Yang, Hao Wan, Hua-Jun Zeng, Zheng Chen, and Wei-Ying Ma. 2005. Support vector machines classification with a very large-scale taxonomy. *ACM SIGKDD Explorations Newsletter*, 7(1):36–43.

Parma Nand, Ramesh Lal, and Rivindu Perera. 2014. A HMM POS Tagger for Micro-Blogging Type Texts. In *Proceedings of the 13th Pacific Rim International Conference on Artificial Intelligence (PRICAI 2014)*.

B Sri Nandhini and JI Sheeba. 2015. Online social network bullying detection using intelligence techniques. *Procedia Computer Science*, 45:485–492.

Andrew Ortony, Gerald L Clore, and Mark A Foss. 1987. The referential structure of the affective lexicon. *Cognitive science*, 11(3):341–364.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. *UT Faculty/Researcher Works*.

Soujanya Poria, Erik Cambria, Gregoire Winterstein, and Guang-Bin Huang. 2014. Sentic patterns: Dependency-based rules for concept-level sentiment analysis. *Knowledge-Based Systems*, 69:45–63.

Michal Ptaszynski, Pawel Dybala, Tatsuaki Matsuba, Fumito Masui, Rafal Rzepka, and Kenji Araki. 2010. Machine learning and affect analysis against cyber-bullying. In *Proceedings of the 36th Annual Convention of the Society for the Study of Artificial Intelligence and the Simulation of Behaviour*, pages 7–16.

David Allister Simanjuntak, Heru Purnomo Ipung, Anto Satriyo Nugroho, et al. 2010. Text classification techniques used to faciliate cyber terrorism investigation. In *Advances in Computing, Control and Telecommunication Technologies (ACT), 2010 Second International Conference on*, pages 198–200. IEEE.

A Squicciarini, S Rajtmajer, Y Liu, and C Griffin. 2015. Identification and characterization of cyberbullying dynamics in an online social network. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pages 280–285. ACM.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Kathleen Van Royen, Karolien Poels, Walter Daelemans, and Heidi Vandebosch. 2015. Automatic monitoring of cyberbullying on social networking sites: From technological feasibility to desirability. *Telematics and Informatics*, 32(1):89–97.

Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012a. Learning from bullying traces in social media. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 656–666. Association for Computational Linguistics.

Jun-Ming Xu, Xiaojin Zhu, and Amy Bellmore. 2012b. Fast learning for sentiment analysis on bullying. In *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining*, page 10. ACM.

M Ybarra. 2010. Trends in technology-based sexual and non-sexual aggression over time and linkages to non-technology aggression. *National Summit on Interpersonal Violence and Abuse Across the Lifespan: Forging a Shared Agenda*.

Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian D Davison, April Kontostathis, and Lynne Edwards. 2009. Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB*, 2:1–7.