Automated Role Detection in Cyberbullying Incidents

Charibeth Cheng and Louie Anson Ng
De La Salle University
2401 Taft Avenue
Manila, Philippines
charibeth.cheng@dlsu.edu.ph

ABSTRACT

Bullying is a problem does not cease even as people age [10]. With the Internet, bullying has also evolved from being a physical experience to a virtual experience, known as cyberbullying. Since traditional bullying involves the participation of different roles, we speculate that the same roles are also present in cyberbullying. These roles include being an accuser, a bully, a defender, a reporter or a victim. By using social media sites as sources for model training, this paper presents a support vector machine- (or SVM-) based model for detecting bullying incidents and the roles they played. The technique used word features such as n-grams, TF-IDF scores, and a list of profane word, combined with the use a feature weighing scheme. The optimal model produced an accuracy of 59.7% using 171 unique word features, with a Kappa statistic 42.3% in detecting the bullying roles; while detecting the bully role produced an accuracy of 80.9% with a Kappa statistic of 55.52%.

CCS Concepts

• Applied Computing□Law, social and behavioral sciences • Computing methodologies□Machine learning.

Keywords

Machine learning, cyberbullying, role detection

1. INTRODUCTION

While social networking sites are primarily used for communication and interaction between peers, there are some who use these technologies to harm other people emotionally. This is called cyberbullying. Cyberbullying is defined as a willful and repeated harm indicted through the use of computers, cellphones, and other electronic devices [9]. A survey conducted in 2012, included 18,687 citizens in 24 different countries around the world, showed that one in ten parents online say their child has experienced cyberbullying, while one in four say they know a child in their community who has experienced cyberbullying[7]. According to another survey in the US, with a sample size of 935 teens aged 12 to 17 years old, teens who use social networking sites like MySpace and Facebook and teens who use the Internet daily were more likely to say that they have been cyberbullied, with 4 in 10 social networking users (39%) having been cyberbullied in some way on a social networking site compared to 22% who do not use social networks[11].

According to a recent survey, at least 8 out of 10 online Filipinos said they accessed social networking sites in the past [20]. With more Filipinos spending longer times online, cyberbullying in the Philippines has become rampant in social networking sites. Several celebrities have been targets of cyberbullying. The following are just some of the notable ones:

- Singer Charice [3] claimed she was cyberbullied for being a closet lesbian, which led to her coming out in June 2013[8];
- Actress Sharon Cuneta closed her Twitter account, after receiving offensive comments on Twitter[21]
- Senate Majority Leader Vicente Sotto claimed to have been cyberbullied, after it was found out that he plagiarized several speeches without citing his sources[1].
- Jamie Paula Salvoza got her monicker *Amalayer* after a commuter posted a video of her raising her voice at a lady security guard inside the LRT station[5].
- Christopher Lao was cyberbullied after he was featured on a news report for driving towards a flooded street and complaining that he was not informed about it [2].

These reported cyberbullying incidents only represent a portion of the numerous cases that have not been reported. This led to the introduction of Republic Act 10175 also known as the *Anti-Cybercrime Act of 2012*, which recognizes cyberbullying as a kind of cybercrime and provides provisions on the consequences for cyberbullying[19]. However, despite this law, it is simply not possible to monitor the Internet constantly for any acts of cybercrime being committed. This is why automated systems must be developed to help monitor activities occurring in cyberspace.

Research had been conducted focusing on the detection of cyberbullying. [4] discussed the importance of a person's gender, seeing how studies show the difference in the way males and females bully each other. On the other hand, [23] analyzed the presence of roles in cyberbullying episodes based on the author posts, as they may be reporting a cyberbullying incident, accusing someone of cyberbullying, revealing himself being bullied, or is a bully himself. [23] says automated cyberbullying detection is important because intervention methods in cyberbullying vary depending on the role played by each participant, and may be helpful in observing how each participant's role changes over time.

SAMPLE: Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

2. ROLES IN CYBERBULLYING

A series of studies were conducted in Finland [13,14,16,17,18] on a participant's role in the traditional bullying process. The roles that were involved in traditional bullying include the following:

- bully
- victim
- assistants of the bully
- reinforcers
- outsiders
- defenders

A bullying scenario starts with the *bully* and the *victim*. Some may start joining in the bullying because they find it fun and hence act as *assistants of the bully*. Others may only give positive feedback to the bully by giving encouraging gestures to the bully or reinforcing the bully by being a part of the audience, thus they are called the *reinforcers*. Then there are those who stay neutral throughout the entire episode, but this does not mean that they are not involved. For reasons that may include indifference or the fear that they may be targeted instead, they allow the bullying to continue without doing anything to stop it. They are called the *outsiders*. The last group is the *defender* who actively protects the person being bullied and tries to do something to make the bullying stop[15].

These roles were also used by [24] in cyberbullying role detection, along with some minor modifications to the existing roles. Two new roles were added to augment the traditional role system after observing the data they collected. The *reporter* merely posts social media about cyberbullying scenarios that he witnessed or knew, while the *accuser* targets the bully and accuses him of bullying without necessarily defending the victim. Only the most frequent roles that appeared in cyberbullying instances were chosen, and these were the *bully*, the *victim*, the *reporter*, and the *accuser*. All the remaining roles were merged into a generic label *Other*.

3. THE CORPUS

We gathered data from YouTube and Facebook. A total of 6000 comments/posts were obtained (1500 for YouTube and 4500 for Facebook) and classified into one of the following roles:

- Accuser
- Bully
- Defender
- Reporter
- Victim
- N/A (No applicable role)

The YouTube dataset contains comments from videos on controversial events in the Philippines that happened in 2013 that were considered controversial[1] as these topics would most likely to encourage users or viewer to post their opinions about the topic. In YouTube, videos that contained at least 100 comments written in Filipino or English or a combination of both, were considered controversial. In Facebook, articles regarding events (politics, showbiz) which contained more than 500 comments written in Filipino or English (or a combination of both) were selected and crawled for comments. The articles were gathered from fan pages of local news channels and newspapers. These pages have a following and often posted events, which attract public opinion.

The dataset was cleaned in order to remove unnecessary symbols present in text. Listed were the most common ones:

- several consecutive periods or periods separated by a space(ex. *mas gwapo sa kanan....ayyyy makukulong din ako*.)
- consecutive or a mixed? and ! (ex. *waaaaa, ang babaw naman ng kaligayahan ninyo, saan ba nakakatawa dun?????????????*)
- hashtags # (ex. *Yoko mag comment #mahirapnabakamakasuhanpa*). Only the # symbol is removed.

These symbols were removed as so they will not be extracted as features later on. Furthermore, the dataset underwent normalization using NormApi [12], to convert shortcut text into their normal forms. Words like *n22log* would be changed to its normal form *natutulog* so that spelling variations of writing a single word will still be counted as one during feature extraction.

Three annotators labeled the roles in the dataset. Each annotator labeled one-third of the dataset. The dataset was randomly split and the source article was not included. Annotators had no idea what topic is being discussed and simply tagged each instance based on how they were perceived it. The following guidelines were given:

- Bully – instances that contain offensive comments towards another.
- Accuser – instances that have some degree of offensiveness and are also targeted, but does not involve the use of profanity
- Defender - instances with positive sentiments like encouragement or is defending someone
- Reporter - instances that relay cyberbullying information, but do not target any specific person
- Victim - instances with negative sentiment, including towards oneself.
- N/A - instances do not belong to any of the 5 classes above.

Table 1: Distribution table for dataset instances

Role	Count
Accuser	621
Bully	1355
Defender	1413
Reporter	188
Victim	78
N/A	2347

Table 1 shows the distribution of instances per role. Non-cyberbullying data was most prevalent, followed by defenders, bullies, accusers, and reporters. The role of the victim had the least number of instances. Further showed that low number of victims was because the targets of bullying are the persons involved in the articles being discussed, so the comments are about them, but the victims do not necessarily make comments.

4. FEATURES USED

We used the following determining the features:

[1] a dispute, argument, or debate, esp. one concerning a matter about which there is strong disagreement and esp. one carried on in public or in the press (Dictionary.com)

1. bag-of-word
2. TF-IDF was performed to determine the most relevant and discriminating words for each role class to determine the words that are most relevant and discriminating. Bigrams and trigrams of the instances were also evaluated for the TF-IDF scores. Stop words were removed prior to listing the TF-IDF
3. presence of profane words, including *anak ng, asshole, bopols, bullshit, bwakangina, cunt, fackers, fak, fuck, gago, ina, inang, inutil, kupal, motherfucker, packers, pi, pota, pucha, pukingina, punyeta, puta, putang, putangina, shit, syete, tang, tangina, tarantado, teteng*.
4. word shape. Instances completely in upper case are considered offensive.

Several experiments were done to identify the final feature set, and this is discussed in Section 6). The final feature set is resulted from the experiment discussed in Section 6.5. For each role class, the top 50 n-grams (based on TF-IDF) were selected to be included in the feature set. The n-grams were combined to form the initial feature set, and then duplicates were removed. A total of 171 word features used to build the classifying model. The combined list from the top n-grams per class in shown in Table 2. The second column indicates if the n-gram applies to a specific role class.

Feature Weighting was done to distinguish common n-grams that appeared in more than 1 role class. The value of each word feature varies depending on which class the instance is labeled as. If the word feature does not belong to the class where it is being labeled as then it assumes the value of 1, otherwise its value is the weight assigned to it for that particular class.

5. BUILDING THE CYBERBULLYING DETECTOR AND ROLE CLASSIFIER

To be able to analyze each role more carefully, each experiment was conducted 7 times, each with a different set of role classes to see if some classes are related in terms of similarity or if they will prove to be difficult to classify. The combination of role classes is as follows:

Set 1: All Classses

This set serves as the baseline for all test results as this aims to be able to determine how each class performs.

Set 2: Bully and N/A

This set only contains the most important roles, which may either be a bully or not have any role at all that indicates that no bullying occurred. This also serves as a reference for comparison against other literature to determine how the use of word features that identify roles can be used to detect cyberbullying.

Set 3: Accuser, Bully, and N/A

The accuser is most relevant to the bully, so they were put together in order to see if the model would be able to differentiate between these two classes.

Table 2. Selected Word Features based on TF-IDF scores

n-gram	Role class[2]
alam, amalayer, anak, aralan, bad, bansa, batas, bayan, buhay, comment, country, dami, diyan/dyan, doon, eat, freedom, go, gusto, huwag, ikulong, joke, paso know, let, libel, like, marami, much, ngayon, people, pera, post, problems, public, rape, sana, serial, speech, stupid, mama, tango, think, time, totoo, ugali, walang,	All
muna, baka, puro, unahin, kailangan, believe, kapal, mayaman, gumawa, problema	Acc
baboy, ina, gago, putang, ina mo, bobo, tang, tanga, fuck, hayop, bakla, bitch, kabayo, hoy, kasama, ulo, bastos, pangit, babae, ganda, mouth, mamatay, akala, anal, gangrape	Bull
do not, justice, love, siguro, way, good, social, hustiya	Def
kaya, video, cctv, case, isip, black, masama, tanong, galling, project, kanya, pinatay, karapatan, parang, interview	Rep
Family, lungkot, mahirap, walang, come, want, side, manahimik, pray, rest, sakit, oh, hard, make, maawa, message, rest in, abroad, going, real, text, sumali, ano ba, peace, job, express	Vic
Haha, best, well, mistake, gawin, lobby, issue, act, gang, ha, sabi, naku, okay, lumabas, rude, tao	N/A

Set 4: Bully, Defender, and N/A

The defender depicts a positive sentiment as compared to the bully that is considered a negative sentiment, so it should be able to provide higher prediction results than the third set.

Set 5: Bully, N/A, Reporter, and Victim

The reporter and victim roles have the lowest number of instances in the dataset, so this test set can show if these two roles are still relevant to the research.

Set 6: Accuser, Bully, Defender, Reporter, and Victim

This set only includes roles with cyberbullying involved to determine how the accuracy will be if the outlier class was removed. Since non-cyberbullying instances make up a third of the dataset, this test set will provide a more realistic result.

Set 7: Accuser, Bully, and Defender

This set contains only the roles with the most number of instances disregarding non-cyberbullying instances, so the victim and reporters were removed instead.

Several experiments were performed to identify the optimal feature set as well as the learning algorithm to be used. The supervised learning algorithms applied to the role classification include SVM, Naïve Bayes and J48, with 10-fold cross validation. The Section 4 discussed the feature set that produced the best results.

6. EXPERIMENTS DONE

6.1 Baseline Results

The first set of experiments involved an initial set of 25 word features per class as well as checking for the presence of any word

[2] Acc for Accuser, Bull for Bully, Def for Defender, Rep for Reporter, and Vic for Victim.

written in all caps and checking for the presence of profane words. However, upon inspecting each set of word features from each class, some word features present in a class were also present in other classes. To see the relevance of these intersecting words, they were initially included as part of the feature set, but any succeeding occurrences were disregarded. From a total number of 150 word features (25 word features multiplied by 6 classes) only unique words were selected and thus was reduced to 93 unique word features.

Table 3: Results using profanity, full capitalization, and classes may have common word features

Test set	Pre	Re	F-Meas	Kappa Stat
All Classes	45.4%	50%	44.7%	24.73%
Bull + N/A	73.4%	72.8%	70%	34.6%
Acc +Bull + N/A	60.1%	62.6%	55.3%	26.65%
Bull + Def + N/A	59.9%	58.6%	56.3%	31.16%
Bull + N/A + Rep + Vic	68.9%	68.2%	63.6%	30.24%
All except N/A	45.7%	52.5%	45.9%	24.35%
Acc + Bull + Def	52.6%	56.6%	51.7%	27.44%

Table 4: Confusion matrix results for Table 3.

	Acc	Bull	Def	N/A	Rep	Vic
Acc	12	83	114	412	0	0
Bull	14	512	97	732	0	0
Def	20	137	527	727	2	0
N/A	17	175	207	1946	1	1
Rep	3	31	52	102	0	0
Vic	3	7	24	42	0	0

The results of the model from Table 3 show that the training sets with all classes joined together perform the poorest, with only 45.4% accuracy when non-cyberbullying data was removed and only 45.7% when non-cyberbullying data was included. On the other hand, the training set with only bullying and non-bullying[3] data garnered the highest results with an accuracy of 73.4%.

The confusion matrix generated after testing the model using ten-fold cross validation shown in Table 4 also shows the seeming difficulty in detecting accuser roles as only 12 out of a total of 600 were correctly classified, with the rest being classified as non-bullying instances. Even with the non-bullying role removed, the model still failed to predict accuser roles, and instead predicts them as other roles. The most disappointing result was that more than half of bully instances were also misclassified as non-bullying instances, as this should not have been the case given the good word features for the bully role. Because of this, the word feature list was revised and more experiments were conducted in order to see if the accuracy could be improved further.

6.2 Removing duplicate n-grams features

For the second set of experiments, intersecting words and other added features were removed instead. Since these words were present in several classes, it could be possible that the intersecting word features caused the labels to become ambiguous and made it harder the model to predict each instance correctly. In this experiment, if a word feature was found to be present in more than 1 class, then it would not be included in the feature set. The original 150 word features were reduced to only 63. The results can be found in Table 5.

Table 5: Results using profanity, full capitalization, and no common word features across classes

Test set	Pre	Re	F-Meas	Kappa Stat
All Classes	44.8%	44.8%	48.4%	21.54%
Bull + N/A	74.4%	74.4%	72.9%	33.8%
Acc +Bull + N/A	58.2%	62%	55.3%	23.8%
Bull + Def + N/A	57.5%	56.5%	53.5%	27.09%
Bull + N/A + Rep + Vic	65.8%	67.8%	62.5%	28.28%
All except N/A	42.1%	49%	42.3%	18.75%
Acc + Bull + Def	49.9%	53%	47.7%	25.5%

Table 6: Confusion matrix results for Table 5

	Acc	Bull	Def	N/A	Rep	Vic
Acc	12	70	96	442	0	1
Bull	14	446	105	770	0	0
Def	10	133	425	845	0	0
N/A	13	139	193	2002	0	0
Rep	4	18	44	122	0	0
Vic	1	5	25	45	0	0

The results in Table 5 show a decrease in all measures. However, for the test set involving only bullying and non-bullying instances, the model scored higher than the results in Table 3 in terms of precision, recall, and F-measure but had a lower Kappa Statistic with a 0.8% difference.

The results can be seen from the confusion matrix for cross validation wherein more bullying instances were misclassified as non-bullying instances due to non-discriminating words that were included as word features. The word features crossed out from the first experiment were word features that also occurred in at least 2 classes, so it did not help improve the accuracy for non-bullying instances. The confusion matrix showed that removing intersecting features did not help the model in classifying the instances correctly. More instances were mislabeled as non-bullying instances, and this may be due to the intersecting words being correlated to the discriminating words for each dataset.

[3] The non-bullying class shall henceforth be identified as the N/A class.

6.3 Removing profanity presence and all caps word presence

The next experiment validates how profanity presence and full capitalization affect model accuracy. In this experiment, both all caps word presence and profanity presence were removed as features.

Table 7 shows very minimal decrease in accuracy, compared to Table 5. This shows that the presence of profanity and words in all caps did not improve accuracy contrary to related literature but decreased it instead. However, this is because of the profanity list that only contains common swear words. Many instances with bully role contained words considered offensive, but not not necessarily swear or curse words.. For example, Filipino words *tanga* and *bobo*, which both mean stupid, and may not be considered profane but are offensive. This might be the reason why profanity presence did not help in improving model accuracy. In the case of full capitalization presence, full capitalization is not only used for "shouting" text, but also on acronyms or simply as a user preference. This suggests that simply checking if a comment has any word written in full capitalization will not suffice.

Table 7: Results using word features only

Test set	Pre	Re	F-Meas	Kappa Stat
All Classes	45.5%	50%	44.8%	24.82%
Bull + N/A	73.4%	72.8%	70%	34.59%
Acc +Bull + N/A	59.9%	62.5%	55.3%	26.57%
Bull + Def + N/A	59.8%	58.5%	56.3%	31.10%
Bull + N/A + Rep + Vic	65.9%	68.2%	63.6%	30.24%
All except N/A	46%	52.5%	45.9%	24.53%
Acc + Bull + Def	52.9%	56.8%	51.7%	27.8%

Table 8: Confusion matrix results for Table 7

	Acc	Bull	Def	N/A	Rep	Vic
Acc	12	80	119	412	0	0
Bull	13	514	97	731	0	0
Def	22	132	524	733	2	0
N/A	18	169	211	1948	0	1
Rep	3	32	51	102	0	0
Vic	3	7	24	442	0	0

By removing the profanity list and full capitalization presence, the model was able to predict more bully and defender roles, although this only gave an incremental difference in accuracy results in comparison to the baseline experiment test. However, this is not enough as the major roles such as the accuser, bully, and defender are still being misclassified as non-bullying instances. There must be some way to be able to include the intersecting words and still be able to differentiate them between classes. For word features that were present in several classes, priority was given to the class wherein it had a higher TF-IDF score. This was done by using weighted values instead of binary values for the word features.

6.4 Weighted Features

In order to distinguish the top common n-grams associated with more than one role class, a weighting system was used that assigns weights to word features in order for the model to be able to distinguish which classes these word features belong to. The value of each word feature varies depending on which class the instance is labeled as. If the word feature does not belong to the class where it is being labeled as then it assumes the value of 1, otherwise its value is the weight assigned to it for that particular class. Table 3 shows the n-gram features used for this experiment and the results of the experiment are shown in Table 9.

The results show a significant improvement as compared to the previous experiments. The consolidated training set that includes all classes showed a 10% increase in precision and a 13% increase in Kappa. This proves that the assignment of weights helps the classifier identify better which word features have a higher precedence in each of the classes. The bullying and non-bullying training set in Table 9 still shows the highest results, with accuracy reaching as high as 80%.

Table 9: Results using weighted features

Test set	Pre	Re	F-Meas	Kappa Stat
All Classes	55.6%	57.7%	53.3%	37.6%
Bull + N/A	80%	79.7%	78.7%	53.17%
Acc +Bull + N/A	68.1%	69.6%	66.2%	42.64%
Bull + Def + N/A	67.7%	66.9%	69.5%	46.06%
Bull + N/A + Rep + Vic	75.9%	75.9%	73.6%	49.84%
All except N/A	53.1%	55.6%	50.2%	29.62%
Acc + Bull + Def	57.7%	60.2%	56.4%	33.89%

Table 5.8: Confusion matrix results for Table 9

	Acc	Bull	Def	N/A	Rep	Vic
Acc	40	131	136	314	0	0
Bull	30	705	104	516	0	0
Def	26	185	664	553	2	0
N/A	12	142	128	2064	0	1
Rep	13	42	62	68	3	0
Vic	4	13	23	31	0	5

6.5 Adding More N-gram Features

The next experiment involves adding more features to the current word feature set. It may be possible that the reason for the low accuracy is because there are too many similar features and not enough discriminating features. This can be seen with the accuser role wherein more instances were labeled as bullies instead of accusers. During this experiment, some word features were replaced with better ones in an attempt to increase model accuracy. Proper nouns, nouns that were considered domain specific, and more common words were removed in order to retain the best features. From the initial list of 25 word features per class (Section 6.1) the number was increased to 50 word features per class. Based on the good results shown in Section 6.2, we removed the duplicate n-grams. A total of 171 word features were obtained. The final list of n-gram features is shown in Table 3.

Table 11: Result using weighted features and more word features

Test set	Pre	Re	F-Meas	Kappa Stat
All Classes	59.7%	60.6%	57.5%	42.3%
Bull + N/A	80.9%	80.6%	79.7%	55.52%
Acc +Bull + N/A	71.7%	72.1%	69.9%	48.03%
Bull + Def + N/A	71.7%	70.4%	69.5%	46.06%
Bull + N/A + Rep + Vic	77%	76.8%	75.1%	51.97%
All except N/A	57%	59.6%	55.7%	36.95%
Acc + Bull + Def	60.4%	63.6%	60.7%	40.1%

Table 11 shows a higher accuracy than the previous experiment, with the exception of the bullying and non-bullying data showing only a 0.9% increase in precision. This is because the increase in features applied to the other classes while the word features of both the bully and non-bullying classes in the previous experiment were already considered enough to generalize the entire class.

Table 12: Confusion matrix results for Table 11

	Acc	Bull	Def	N/A	Rep	Vic
Acc	70	76	165	306	0	0
Bull	33	750	86	485	1	0
Def	67	93	734	509	10	0
N/A	22	134	148	2040	2	1
Rep	15	12	66	73	22	0
Vic	4	5	21	35	0	10

With more relevant features being added, the accuser, bully, and defender roles were able to have more correctly classified instances as seen in the experiment confusion matrix for 10-fold cross validation. However, the addition of word features meant that more irrelevant features were also being added to the feature set, resulting in less non-bullying instances. Improvement can also be seen in reporter and victim instances although this increase is minimal.

6.6 Learning Algorithm Comparison

Table 13. Comparison of Learning Algorithms Used

Algorithm	Pre	Re	F-Measure	Kappa Stat
SVM	59.7%	60.6%	57.5%	42.3%
J48	43.8%	50.6%	45.8%	22.54%
Naive Bayes	53.2%	54.9%	52.4%	34.53%

Finally, the experiment that gave the highest test results (Section 6.5) was also executed using different algorithms to see how it fared against other known supervised learning algorithms. Based from related literature, Naive Bayes classifiers and decision trees were tested alongside support vector machine as a form of comparison.

Table13 shows that SVM had the highest accuracy among the 3 most used algorithms in supervised learning. Naive Bayes classifier only assumes that each feature is independent from the other to make a classification. However, based from the results of the experiments in Table 5.3 and Table 5.1 involving the removal of intersecting words that show a negative effect in accuracy, it can be seen that some words are dependent on one another in order to be able to produce good results. In the case of J48, it performed poorly than support vector machines because the intersection of word features made it harder to classify each instance to its class. This comparison shows that SVMs are best suited for this type of classification as supported by researches done by [4,6,24].

7. Conclusion

The research detected cyberbullying roles present in texts obtained through social media. These roles include the accuser, bully, defender, reporter and victim. Data used for the research was collected from social media sites Facebook and YouTube.

The main features used to create the model in this research were word features derived from the dataset using Term Frequency Inverse Document Frequency (TF-IDF) for determining relevant word features. N-gram language modeling was applied to consider each term in the corpus as a feature, up to a window size of 3 terms per feature. For each class, the top 50 word features with the highest TF-IDF scores were considered. Subsequent duplicates of word features were then removed, and a total of 171 unique word features was obtained. The word features were then assigned weighted values based on the TF-IDF score ranking. The final model was able to achieve an accuracy of 59.7% through 10-fold cross validation.

Even though bigrams and trigrams were included in the TF-IDF ranking, words that had the highest TF-IDF scores were mostly unigrams. This is because the high-ranking bigrams and trigrams contained stop words, which were removed prior to computing TF-IDF scores.

The Bully role is seen to be the most distinguishable of all roles. This is attributed to the fact that the n-grams associated to the Bully class are indeed mostly profane, offensive, and derogatory. The Bully role also has the least number of top n-grams appearing in the other roles. The second most distinguishable role is the Defender. On the other hand, the *accuser* and *reporter* roles are most difficult to classify. In fact, in all the experiments performed, these two roles are usually misclassified. Thus, the results for classification that involved the accuser or reporter roles always have the lowest results. Again, this result may be attributed to the top n-grams associated with these roles. The words specific to the accuser role and the reporter role do not represent who they are, as compared the n-grams associated with the Bully, Defender and Victim roles. Almost all the words associated with the Bully role are offensive. The words related to the Victim deal with sadness and cries for help. For the Defender role, although the list is short, the words are related to reasons why bullying should stop.

On all the experiments conducted, the test set with only bullying and non-bullying data performed consistently well. While those that involved the accuser and reporter role, performed worst.

The experiments showed that full-text capitalization as a feature has little effect on the prediction. Likewise, profanity presence also contributed little to the prediction accuracy. This is because our profane word list only include swear and curse words, but lack derogatory and offensive words.

Though the technique used in this research may be applied to other languages, the profane word list will vary per language.

8. FUTURE WORK

Despite the high amount of labeled instances, the model failed to classify the accuser roles effectively. In order to be able to determine which roles are present in the local text, more analysis must be made by browsing through more data. The accuser and bully roles seem to be similar to each other, while the reporter and victim roles may not be present at all. Having an understanding of the way Filipinos write their sentiments online will help in identifying, which roles are more related to localized cyberbullying.

Though the use of a profanity list did not improve the accuracy of the current model, this may be because the words the model was supposed to look for were not present in the profanity list containing only 30 words. An offensive word list should be used instead in this scenario, and offensive words can be categorized according to a level of severity. For example, the word *tangina* can be considered very offensive, while the word *tanga* is not as harsh as the previous. By assessing the offensiveness of a post based on this, accusers and bullies may be differentiated. Aside from this, the use of the words in the uppercase as features can be modified as a weighted feature to see if it will still decrease model accuracy. If a comment contains more words were written in the uppercase then it should convey stronger emotions than one with a few or no words that are written in the uppercase.

The model created was found to be language independent. In order for the model to specifically target the Filipino language, a knowledge base must be built. This may be in the form of Filipino word etymologies, orthographies, and ontologies. This information should be available in order to be able to extract more features from Filipino text that cannot be applied to other languages.

Word features were used in creating the feature set for classification. The advantage of using bag-of-words approach is that it only contains relevant information about the classes. Words that frequently occur in a class are bound to have some form of association with that class. However, the problem with this is that it does not consider the context in which the word was used. An example would be the following comment: *where is the justice in this?? shame on you Binay!!*

The instance was annotated with the accuser role, however since the word *justice* had a high weighted value for the defender class, it was misclassified as a defender instead. If the word *shame* did not have a higher value than the word *justice* it would not be classified correctly. By determining how each term was used in the comment, the model should be able to correctly assign the role given an instance.

Although NormAPI[12] was able to correct some of the word shortenings in the dataset, it can be improved further in order to detect more complex variations of short-cut text. The ability of the normalizer to correct words is limited by its dictionary, so a self-learning normalizer that would automatically add new words to its dictionary would be very beneficial in this case.

9. REFERENCES

[1] Ager, M. 2012. I'm a victim of cyber-bullying-Sotto. *Inquirer.net.* Retrieved from http://newsinfo.inquirer.net/260062/im-a-victim-of-cyber-bullying-cries-sotto-as-he-calls-for-scrapping-of-plagiarized-part-of-speech-from-records. (August 2012)

[2] Allego, F. 2011. Christopher Lao is Victim of Cyberbullying. *Blogger Engineer.* Retrieved from http://bloggerengineer.com/2011/08/christopher-lao-is-victim-of-cyber-bullying.html. (August 2011)

[3] Buan-Deveza, R. 2011. Charice admits to being a victim of cyber-bullying. *abs-cbnNews.com*. Retrieved from http://www.abs-cbnnews.com/entertainment/06/21/11/charice-admits-being-victim-cyber-bullying. (June 2011)

[4] Dadvar, M., & de Jong, F. 2012. Cyberbullying detection: A step toward a safer Internet yard. *In Proceedings of the 21st international conference companion on World Wide Web*, 121-126.

[5] dela Cruz, K. 2012. 'Amalayer' becomes buzzword among netizens. *abscbnNews.com*. Retrieved from http://www.abs-cbnnews.com/lifestyle/11/14/12/amalayer-becomes-buzzword-among-netizens. (November 2012)

[6] Dinakar, K., Reichart, R., & Lieberman, H. 2011. Modeling the detection of textual cyber-bullying. *In Proceedings of 2011 International AAAI Conference on Weblogs and Social Media.*

[7] Gottfried, K. 2012. One in Ten (12%) Parents Online, Around the World Say Their Child Has Been Cyberbullied, 26% Say They Know of a Child Who Has Experienced Same in Their Community. *IPSOS*. Retrieved from http://www.ipsos-na.com/news-polls/pressrelease.aspx?id=5462. (January 2012)

[8] Gruttadaro, A. 2013. Charice comes out: 'Yes, I'm a lesbian'. *Hollywood Life.* Retrieved from http://hollywoodlife.com/2013/06/02/charice-comes-out-lesbian-gay-interview. (June 2013)

[9] Hinduja, S., & Patchin, J. 2009. *Bullying beyond the schoolyard: Preventing and responding to cyberbullying.* Corwin Press.

[10] Krantz, L. 2012. No age limit in bullying. *Metro West Daily News*. Retrieved from http://www.metrowestdailynews.com/news/x345285450/No-age-limit-on-bullying. (July 2012).

[11] Lenhart, A. 2007. Cyberbullying is not worse than physical bullying. *Pew Research Center*. Retrieved from http://www.pewinternet.org/Reports/2007/Cyberbullying.aspx. (June 2007)

[12] Nocon, N., Cuevas, G., Magat, D., Suministrado, P., & Cheng, C. 2014. Normapi: An API for normalizing Filipino shortcut texts. *In Proceedings of International Conference on Asian Language Processing 2014 (IALP 2014)*. 207-210.

[13] Salmivalli, C. 1998. Intelligent, attractive, well-behaving, unhappy: The structure of adolescents' self-concept and its relations to their social behavior. *Journal of Research on Adolescence*, 8, 333-354.

[14] Salmivalli, C. 1998. *Not only bullies and victims. Participation in harassment in school classes: Some social*

and personality factors. Unpublished doctoral dissertation, Annales Universitatis Turkuensis.

[15] Salmivalli, C. 1999. Participant role approach to school bullying: implications for interventions. *Journal of Adolescence*, 22, 453-459.

[16] Salmivalli, C., Huttunen, A., & Lagerspetz, K. 1997. Peer networks and bullying in schools. *Scandinavian Journal of Psychology*, 38 , 305-312.

[17] Salmivalli, C., Lagerspetz, K., Björkqvist, K., Österman, K., & Kaukiainen, A. (1996). Bullying as a group process: Participant roles and their relations to social status within the class. *Aggressive Behavior*, 22, 1-15.

[18] Salmivalli, C., Lappalainen, M., & Lagerspetz, K. 1998. Stability and change of behavior in connection with bullying in schools. *Aggressive Behavior*, 24, 205-218.

[19] Toral, J. 2012. 17 Cybercrimes Covered Under Cybercrime Prevention Act-Republic Act 10175. *Digital Filipino*. Retrieved from http://digitalfilipino.com/introduction-cybercrime-prevention-act-republic-act-10175. (September 2012)

[20] Tuazon, J. 2011. Pinoys spend 10 hours weekly websurfing at home. *GMA News Online*. Retrieved from http://www.gmanetwork.com/news/story/229165/scitech/pinoys-spend-10-hours-weekly-websurfing-at-home. (August 2011)

[21] Velunta, L. 2012. Sharon Cuneta quits twitter due to cyberbullying. *Philippine Online Chronicles*. Retrieved from http://thepoc.net/index.php/sharon-cuneta-quits-twitter-due-to-cyberbullying/. (March 2012)

[22] Villanueva, M. 2011. 'Think Before You Click!' GMA urges responsible tweeting. *GMA News Online*. Retrieved from http://www.gmanetwork.com/news/story/226420/scitech/think-before-you-click-gma-urges-responsible-tweeting. (July 2011)

[23] Xu, J.-M., Zhu, X., & Bellmore, A. 2012. Fast learning for sentiment analysis on bullying. *In Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining*, 10.

[24] Zhu, J. X., Xu, J.-M., Jun, K.-S., & Bellmore, A. 2012. Learning from bullying traces in social media. *In Proceedings of NAACL HLT '12 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 656-666.