Inappropriate Expressions Recognizer

Ria A. Sagum
Faculty, PUP, Sta. Mesa Manila
Email:rasagum@gmail.com

Joshua S. Dapitan
CCIS, PUP, Sta. Mesa Manila
Email: joshuadapitan@gmail.com

Anjanette R. Lasala
CCIS, PUP, Sta. Mesa Manila
Email: anjlasala@gmail.com

ABSTRACT

Inappropriate expressions recognition is a task of recognizing words whose usage is in an inappropriate context, which is the words used in offensive sense or sexually explicit sense. In this study, the researchers developed a prototype inappropriate expressions recognition model that analyzes sentences in a phrase-level orientation to determine if the word usage is inappropriate or not. Our approach is by using Bootstrapping, Naïve Bayesian Classification, N-Gram Language Modelling, Bag of Words Model, and Hidden Markov Modelling. After the experiment is executed, considering 500 comments, the following rates were computed based from the gathered data: Recall – 66.84%, Precision – 73.12%, Specificity – 96.70% and F-Measure – 69.84%. We were able to conclude that using the definitions as the basis for inappropriate expressions recognition is possible to give possible results and inappropriate expressions can be modeled despite the noise produced by the informal definitions of UrbanDictionary.

Keywords **– Inappropriate Expressions, Bootstrapping, Semi-Supervised Learning**

1. INTRODUCTION

Inappropriate Expressions is one of the problems in a behavioral sense [1]. Inappropriate Expressions mostly causes problems in literary management like cyber bullying, and exposure of children to other textual data that may cause other interests like crime, sex, etc.

The solution is to analyze the inappropriate expressions and model the inappropriate language. The problem is, due to the inherent ambiguity of the language, there is a hard time to recognize the real meaning if whether the expression gets inappropriate, making it harder to be recognized. For example, the word screw may mean the actual screw material, or a slang term for sexual intercourse, which is in accordance to the definition of WordNet of the said example. There are also implemented solutions such as word filters which is by detection of the words, which affects the recognition because of the disregarding of the context.

The objective of the study is to use a machine learning methodology, which is bootstrapping, that models Inappropriate language, in which embodies Inappropriate Expressions. With this solution, users may find inappropriate expressions in textual data, which can be used as a tool for prevention of the exposure of the inappropriate expressions to those who are not concerned. This also models the sentence-level context analysis to identify the inappropriateness of an expression with the use of Lexical Syntactic Features and Grammar Relations as a support to the said computational model that is used to solve the problem of modeling the inappropriate language.

2. RELATED WORKS

The researchers introduced the study of bullying to the NLP Community. Bullying, in both physical and cyber worlds (the latter known as cyberbullying), has been recognized as a serious national health issue among adolescents. One is being bullied or victimized when he or she is exposed repeatedly over time to negative actions on the part of others. There are wide ranges of emotions expressed in bullying traces. After manually inspecting a number of bullying traces in Twitter, our domain experts identified seven most common emotions such as anger, embarrassment, empathy, fear, pride, relief and sadness. Analyzing Social Media to Detect Cyber Bullying using Sentiment Mining found that "sentiment analysis is the task of finding the opinions of people about specific textual entities. The decision making process of people is usually affected by the opinions formed by domain authorities and the proliferation of online discussions [2].

Sentiment Analysis has the potential to identify victims who pose high risk to themselves or others, and to enhance the scientific understanding of bullying overall victims usually experience negative emotions such as depression, anxiety and loneliness. In extreme cases such emotions are more violent or even suicidal. Detecting at risk individuals via sentiment analysis enables potential interventions. In addition, social scientists are interested in sentiment analysis on bullying traces to understand participants' motivations [3].

The Lexical Syntactical Feature (LSF) approach from the research Detecting Offensive Language in Social Media to Protect Adolescent Online Safety is to identify offensive contents in social media, and further predict a user's potentiality to send out offensive contents. It includes two phases of offensive detection. Phase 1 aims to detect the offensiveness on the sentence level and Phase 2 derives offensiveness on the user level. In Phase 1, the researchers apply advanced text mining and natural language processing technique to derive lexical and syntactic features of each sentence. Using these features, we derive an offensive value for each sentence. In Phase 2, we further incorporate user-level features where we leverage research on authorship analysis. The system consists of pre-processing and two major components: sentence offensiveness prediction and user offensiveness estimation. During the pre-processing stage, user's conversation history is chunked into posts, and then into sentences. During sentence offensiveness prediction, each sentence's offensiveness can be derived from two features: its word's offensiveness and the context. The researchers use lexical feature to represent words' offensiveness in a sentence, and syntactic feature to represent context in a sentence. Words' offensiveness nature is measured from two lexicons. For the context, we grammatically parse sentences into dependency sets to capture all dependency types between a word and other words in the same sentence, and mark some of its related words as intensifiers. The intensifiers are effective in detecting whether offensive words are used to describe users or other offensive words. During user offensiveness estimation stage, sentence offensiveness and users' language patterns are helped to predict user's likelihood

of being offensive. Experimental result shows that the LSF sentence offensiveness prediction and user offensiveness estimate algorithms outperform traditional learning based approaches in terms of precision, recall and f-score. It also achieves high processing speed for effective deployment in social media [4].

Very few other research teams are working on the detection of cyber bullying. A misbehavior detection task was offered by the organizers of CAW 2.0. but only one submission was received. It is determined that a baseline text mining system (using bag of words approach) was significantly improved by including sentiment and contextual features. Even with the combined model, a support vector machine learner could only produce a recall level of 61.9% [5].

The researchers of Detecting Offensive Tweets via Topical Feature Discovery over a Large Scale Twitter Corpus in proposed a novel semi-supervised approach for detecting profanity-related offensive content in Twitter. They introduced an approach that exploits linguistic regularities in profane language via statistical topic modeling on a huge Twitter corpus, and detects offensive tweets using these automatically generated features. Their step by step processes are as follows: (a) Bootstrap between twitters and tweets based on a seed word set to obtain training tweets for topic model learning; (b) Topic models are learned via a generative LDA approach; (c) Tweets in a holdout testing set are processed in the same fashion as in (a); (d) Topic distributions are inferred for each testing tweet by the topic model learned in step (b); (e) Seed words are applied against each testing tweet, leading to a binary lexicon feature; (f) ML models are built and evaluated. The keyword matching technique has been shown to perform very well in the literature and achieved a TP of 69.7% with an FP of 3.77% in our experiment. While keeping the FP on the same level as the baseline, our approach had a TP of 75.1% over 4029 testing tweets using Logistic Regression, a significant 5.4% improvement over the baseline [6].

Sentiment classification methods can be categorized into three types: unsupervised [7], supervised [8], and semi-supervised [9]. Compared to supervised and unsupervised methods, semi-supervised methods for sentiment classification become more and more popular due to their making use of both the labeled and unlabeled data. This paper mainly focuses on semi-supervised methods for sentiment classification.

Open Problems in Efficient Semi-supervised PAC Learning address semi-supervised learning for imbalanced sentiment classification. It adopts under-sampling to generate multiple sets of balanced initial training data and then propose a novel semi-supervised learning method based on random subspace generation which dynamically generates various subspaces in the iteration process to guarantee enough variation among the involved classifiers. Evaluation shows that semi-supervised method can successfully make use of the unlabeled data and that dynamic subspace generation significantly outperforms traditional static subspace generation [10].

Word Filtering is one of the most commonly used techniques in the recognition of the inappropriate expressions. Most of this is implemented via a word list and some regular expressions. It is not effective in grasping the context of inappropriateness in the expression causing more expressions to be falsely recognized.

The solutions using Lexical Syntactic features and Grammatical Relations play a big role in the recognition of determining inappropriate expressions. The said features help in identifying the usage in the sentence level, and then how it affects the person related in the expressions, with but lacks independence due to the loss of machine learning techniques. Bootstrapping approach for learning has been proven to be effective in learning, and modeling of linguistic data, though it is preferred to have forms of supervision rather than being unsupervised. There is a need for learning for the features of inappropriate expressions to further model the inappropriate language, in which it contains inappropriate expressions, thus coming up with a bootstrapping methodology.

3. SYSTEM ARCHITECTURE

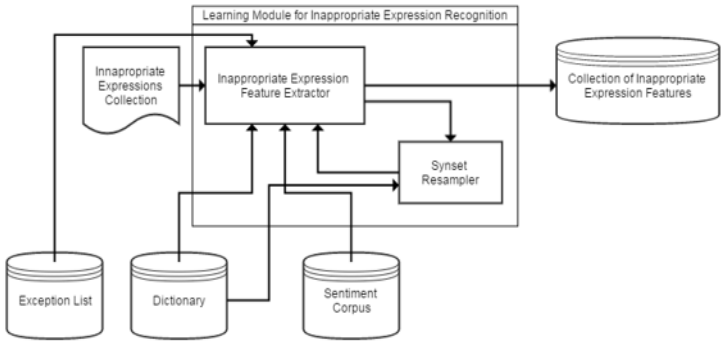

Figure 1 – Learning Module for Inappropriate Expressions

Figure 1 shows the Learning Module for inappropriate expressions. The learning module input consists of a file that contains lists of words that is deemed as Inappropriate. The learning module evaluates the Inappropriateness with basis of its polarity value in the Sentiment Corpus (which will be implemented via SentiWordnet) and the polarity of the definition of the word, in which it will be extracted in two dictionaries (via WordNet[11] dictionary and Urban Dictionary website via Web Scraping) with the implementation of Naïve Bayes model. The Inappropriate expression back propagation will be done by extracting the feature in the definition that made the input inappropriate, and will be collected to the inappropriate expression features knowledge base. The synset resampling gets the synsets of the word and will undergo to the phases undergone by the original word. There is an exception list implemented to compensate and filter the noisy data descriptions of Urban Dictionary that causes false positives. The training module repeats this per word in the collection until all are evaluated and there are no more synsets to be resampled. After the Learning phase, the Learner Module offsets a threshold between on the mean and the global minima of the feature set as a computational borderline for inappropriate expressions. This learning is for the Unigram Expression training.

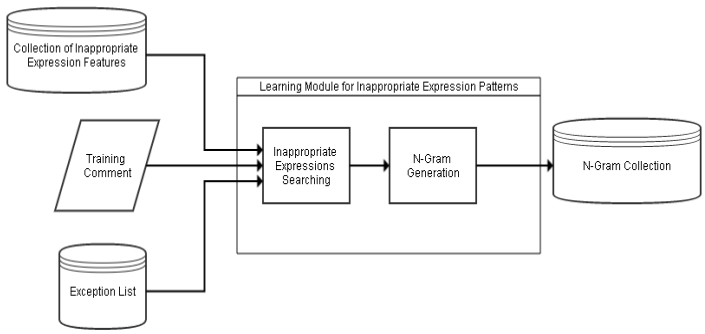

Figure 2 – Learning Module for Inappropriate Expressions Patterns

Figure 2 shows the Learning Module for inappropriate expressions which is used in relation to the neighboring words. This learning module input consists of a training comment and inappropriate expressions feature sets. The learning module finds inappropriate expressions on the input and tags them. There is an exception list implemented to remove the appropriate expressions that are tagged as inappropriate to filter the noisy data descriptions of Urban Dictionary that causes false positives. After tagging the inappropriate expressions, an N-Gram generator generates N-Grams from 2-Grams to 5-Grams based on the POS Tags of the neighboring grams.

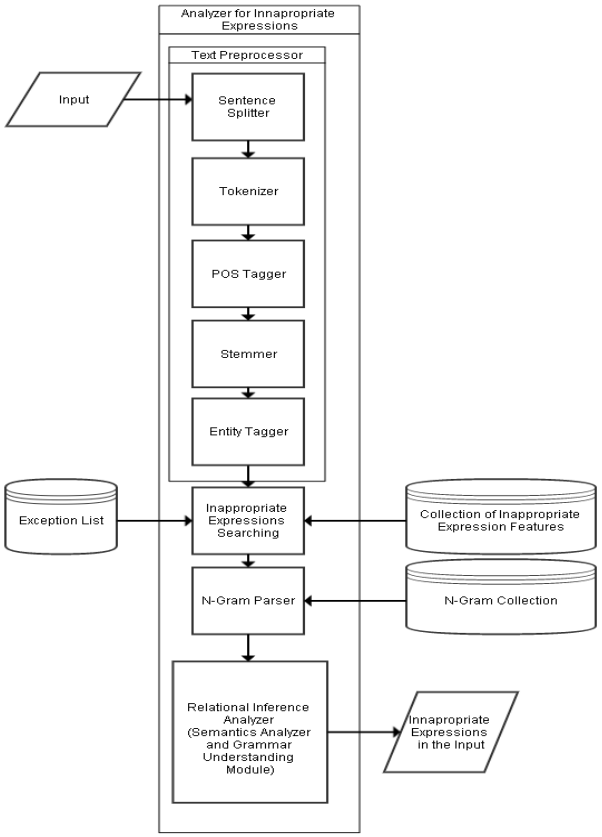

Figure 3 – Analyzer for Inappropriate Expressions

Figure 3 shows the Analyzer Module, there will be an input of a comment. A document will undergo preprocessing. The preprocessing phase consists of Sentence splitting, Tokenization, Part-of-Speech Tagging, and Stemming (for the extraction of base form), and Entity Recognition. Preprocesses are done via Stanford CoreNLP tool [12] and MIT JWI Stemmer [13] for the stemming. After undergoing preprocessing, for each sentence there will be a search for candidates in inappropriate expressions, which will be based on the collected features in the knowledge base. The basis for Inappropriate Expressions searching is the features collected in the definitions in WordNet and UrbanDictionary. There is an exception list implemented to remove the appropriate expressions that are tagged as inappropriate to filter the noisy data descriptions of Urban Dictionary that causes false positives. Then the sentence will undergo to the N-Gram parsing to determine the probable usage of the inappropriate expressions in an inappropriate sense. After the parsing, the Relational Inference Analyzer determines the inappropriateness of the candidate words based each words' Lexical Syntactic Features and its grammar relations to the other existing words in the same input.

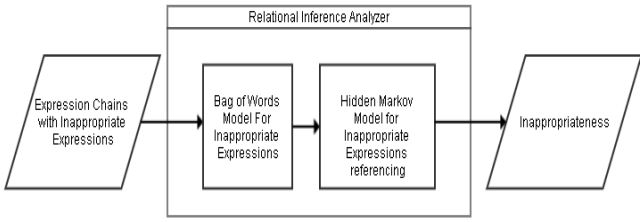

Figure 4 – Relational Inference Analyzer

Figure 4 shows Relational Inference Analyzer is composed of multiple models that determines the Inappropriateness of the expression. First is the Bag of Words model that counts the candidate inappropriate expressions if they are at least 40% of the expression chain. The Hidden Markov Model determines candidate Inappropriate Expressions in relation to other inappropriate expressions and classified entities to determine the inappropriateness.

4. EVALUATION AND RESULTS

The researchers gathered 500 YouTube and 9gag comments through the use of purposive-quota sampling due to its unknown total population. The results were tallied for each tagged inappropriate expression, such as number of inappropriate expressions that must be tagged by the model, number of inappropriate expressions that were correctly tagged by the model based on the expert (TP), number of inappropriate expressions that were correctly tagged by the model but not the expert (FP), number of inappropriate expressions that were correctly tagged by expert but not by the model(FN) and number of appropriate expressions that were correctly tagged by the expert and model (TN). From these classifications of results, the researchers computed for the following equations:

$$\frac{TP}{TP + FP} \quad (1)$$

$$\frac{TP}{TP + FN} \quad (2)$$

$$\frac{2PR}{P + R} \quad (3)$$

$$\frac{TN}{TN + FP} \quad (4)$$

$$\frac{FN + FP}{TP + FP + TN + FN} \quad (5)$$

Equation (1) was used to compute for the Precision. Precision is the percentage of identified expressions that are inappropriate. The table below shows the overall evaluation in terms of Precision.

Table 1 – Overall Evaluation in terms of Precision

CLASSIFICATION	TOTAL NUMBER	PRECISION TP/TP+FP
TP	389	**73.12**
FP	143	

Table 1, presents the overall performance of the model in terms of Precision. Precision is the percentage of identified expressions that are inappropriate. The overall precision of the 50 files tested was 73.12% because according to the researchers, some of the expressions that are contextually appropriate are still recognized as inappropriate in context. The primary reason for such was seen in Figure 5, it is because some of the Lexical Syntactic features of the Expression are noise data from definitions of UrbanDictionary.

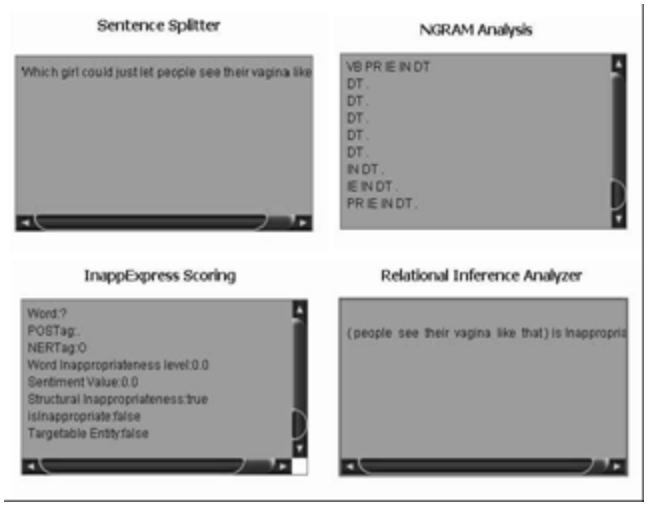

Figure 5 – Sample Comment with False Positive Result

Equation (2) was used to compute for the Recall. Recall is the percentage of inappropriate expressions that are correctly identified. The table below shows the overall evaluation in terms of Recall.

Table 2 – Overall Evaluation in terms of Recall

CLASSIFICATION	TOTAL NUMBER	RECALL TP/TP+FN
TP	389	**66.84**
FN	193	

Table 2, presents the overall performance of the model in terms of Recall. Recall is the percentage of inappropriate expressions that are correctly identified. To compute this, the researchers were accompanied by an English Teacher as their expert in evaluating each comment found in each file. The overall recall of the 50 files tested was 66.84% because according to the researchers, some of the expressions that are contextually inappropriate are still unrecognized by the model. The primary reason for such was seen in Figure 6, it is because grammar relations are not fully established due to the uncontrolled scoping of the phrase-level orientation from the N-Gram Language Model.

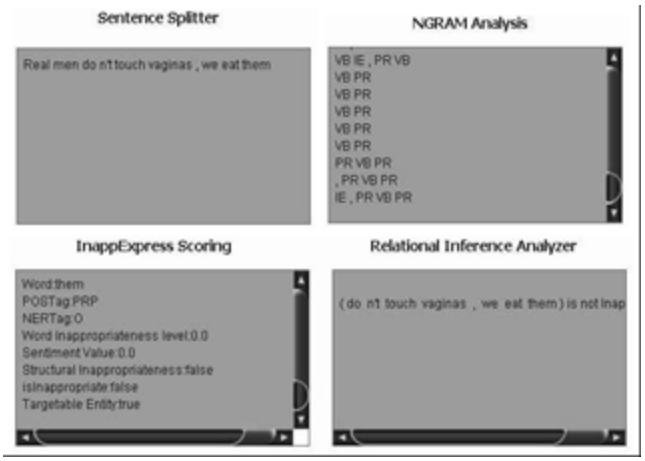

Figure 6 – Sample comment with False Negative Result

Equation (3) was used to compute for the overall performance of Inappropriate Expressions Recognizer. The table below shows the overall performance of the model.

Table 3 – Overall Performance in terms of F-Measure

CLASSIFICATION				PRECISION	RECALL	F-MEASURE (2PR/(P+R))
TP	FP	TN	FN			
389	143	5662	193	73.12	66.84	**69.84**

Table 3, shows the overall performance of Inappropriate Expressions Recognition using Bootstrapping as Semi-Supervised Learning in terms of F-Measure. The average F-Measure of the 50 files tested was 69.84% because based on the previous tables, the recognizer produced low percentage of recall which affected the performance of the model. Hong et. al. study showed in his findings that their study got a recall of 61.9%. It simply implies that this model has a 5% difference in recall compared to Hong's study and that bootstrapping is effective if more training data are to be fed to the model.

Equation (4) was used to compute for the overall performance in terms of Specificity. Specificity is the rate of results without the condition, which the negative test result. The table below shows the overall evaluation of the model in terms of Specificity.

Table 4 – Overall Evaluation in terms of Specificity

CLASSIFICATION	TOTAL NUMBER	SPECIFICITY TN/TN+FP
TN	5662	**97.50**
FP	143	

Table 4, shows the overall performance of Inappropriate Expressions Recognition using Bootstrapping as Semi-Supervised Learning in terms of Specificity. Specificity is the rate of the results without the condition, which has a negative test result. The average Specificity of the 50 files tested was 97.50% because contextually appropriate expressions that were not tagged by both model and

expert were greater than the tagged inappropriate expressions that are not contextually inappropriate to the expert.

Equation (5) was used to compute for the overall performance in terms of Error-Rate. Error-Rate is the percentage of errors encountered during evaluation. The table below shows the overall evaluation of the model in terms of Error-Rate.

Table 5 – Overall Evaluation in terms of Error-Rate

CLASSIFICATION				ERROR RATE
TP	FP	TN	FN	
389	143	5662	193	**5.28**

Table 5, shows the overall evaluation of Inappropriate Expressions Recognition using Bootstrapping as Semi-Supervised Learning in terms of Error Rate. Error Rate is the percentage of errors encountered during the evaluation. The average Error Rate of the 50 files tested was 5.28%. The primary reason why the researchers got that error rate result was that, on the evaluation phase, the researchers got the highest number of true negatives which greatly affects the overall result.

5. CONCLUSION

Based from the findings of the study entitled "Inappropriate Expressions using Bootstrapping as Semi-Supervised Learning" the proponents reached the following conclusions through the series testing and evaluation:

1. There is still an imbalanced classification between appropriate and inappropriate expressions.
2. The Exception List has provided a workaround against the noisy definitions of UrbanDictionary that causes false positives.
3. Bootstrapping has been an effective Semi-Supervised learning schema for Inappropriate Expressions.
4. The performance of the model can be improved if more feature functions were fed into the model.

6. RECOMMENDATIONS

The following suggestions might be helpful for those future researchers who will also specialize in any topic relating to Inappropriate Expressions Recognition:

1. Compare the significant difference of the developed Inappropriate Expressions Recognizer that utilizes different algorithm to further emphasize the usefulness of the Bootstrapping approach.
2. This model can be improved by broadening the feature word neighboring to sentence neighboring in order to determine correctly the context of the detected inappropriate expressions.
3. There is a need for a supervised N-Gram Modeling to model the inappropriate expressions in a proper phrase level scoping.
4. There should be a control of definitions to be used in avoiding noise of false positives and false negatives.
5. Create your own dictionary which is consulted from the expert to avoid imbalanced classification between appropriate and inappropriate expressions.
6. There is a need for a methodology to handle the noisy definitions of UrbanDictionary to remove the need of an exception list.
7. There may be a need of usage of other dictionaries like Merriam-Webster's Dictionary and Oxford Dictionary (Merriam-Webster's Dictionary and Oxford's dictionary needs legal permissions before usage).
8. There may be a need for the recognition of multiple word idiomatic inappropriate expressions.

ACKNOWLEDGEMENTS

This thesis becomes a reality with the kind of support and help of many individuals. We would like to extend our sincere thanks to all of them.

To our parents, for the encouragement, heartening support and understanding they have shown in every task we do.

To our classmates and friends, for extending their encouraging support in the process of constructing this paper even beyond their busy schedules. All the positivity for us as we continue our walk in this difficult yet fulfilling journey.

To Dr. Vivien Domingo, for her constructive criticisms that guided us in appropriately painting the words that this paper contains.

To our Thesis Adviser, Prof. Ria A. Sagum, for imparting her knowledge and expertise in this study and for giving helpful comments and suggestions for the progress of the researchers' paper.

Foremost, we want to offer this endeavor to our GOD Almighty for the wisdom he bestowed upon us, the strength, peace of mind and good health in order to finish this research.

REFERENCES

[1] Jay, T.(2013).*Why We Curse?* John Benjamins Publishing Company, North Adams, Massachusetts.

[2] Dr Y Bi (n.d.). Analysing Social Media to Detect Cyber Bullying using Sentiment Mining. School of Computing and Mathematics, Faculty of Computing and Engineering at the Jordanstown Campus of the University of Ulster: http://www.findaphd.com/search/ProjectDetails.aspx?PJID=56055

[3] Bellmore, A., Xu, J.M. and Zhu, X.(August 2012). Fast Learning for Sentiment Analysis on Bullying. *In Proceedings of ACM 978-1-4503-1543-2/12/08*.

[4] Chen, Y., Xu, H., Zhou, Y. and Sencun, Z.(2012). Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. *In Proceedings of PennState College of Information Sciences and Technology*.

[5] Hong, L., Xue, Z. and Yin, D.(April 2009). Detection of Harassment on Web 2.0. *In Proceedings of Lehigh University: Computer Science and Engineering*.

[6] Fan, B., Hong, J., Rose, C., Wang, L. and Xiang, G.(2012). Detecting Offensive Tweets via Topical Feature Discovery over a Large Scale Twitter Corpus. *In Proceedings of ACM 978-1-4503-1156-4/12/10*.

[7] Turney, P.(2002). Thumbs up or Thumbs down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *In Proceedings of ACL-02*, pp. 417-424.

[8] Lee, L., Pang, B. and Vaithyanathan S.(2002). Thumbs Up? Sentiment Classification usingMachine Learning Techniques. *In Proceedings of EMNLP-02*, pp. 79-86.

[9] Melville, P. and Sindhwani, V.(2008). Document-Word Co-regularization for Semi Supervised Sentiment Analysis. *In Proceedings of ICDM-08*, pp. 1025-1030.

[10] Balcan, M. and Blum, M. (n.d.). Open Problems in Efficient Semi-Supervised. *InProceedings of National Science Foundation Grant CCF-0514922 and a Google Research Grant*.

[11] George A. Miller (1995). WordNet: A Lexical Database for English.
Communications of the ACM Vol. 38, No. 11: 39-41.

[12] Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McCloskv. 2014. The Stanford CoreNLP Natural Language Processing Toolkit In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55-60.

[13] Finlayson, Mark Alan (2014) Java Libraries for Accessing the Princeton Wordnet: Comparison and Evaluation. In H. Orav, C. Fellbaum, & P. Vossen (Eds.), *Proceedings of the 7th International Global WordNet Conference* (GWC 2014) (pp. 78-85). Tartu, Estonia.

Ria A. Sagum was born in Laguna, Philippines on August 31, 1969. She took up Bachelor in Computer Data Processing Management from the Polytechnic University of the Philippines and Professional Education in Eulogio Amang Rodriguez Institute of Science and Technology. She received her master degree in De La Salle University in 2012. She is currently an instructor in both Polytechnic University of the Philippines in Sta. Mesa, Manila and University of Santo Tomas in Manila. Prof. Sagum has been a presenter of different conferences, including the 2012 International Conference on e-Commerce, e-Administration, e-Society, e-Education, and e-Technology and is a member of the Computing Society of the Philippines and the Natural Language Processing Special Interest Group.

Joshua S. Dapitan is currently pursuing his Bachelor of Science in Computer Science at Polytechnic University of the Philippines, Mabini Campus. Has knowledge in different programming languages such as C, Java, C#, Python, MATLAB, PHP, Javascript, and SQL. He is interested in Data Mining, Machine Learning, Algorithm Analysis, Artificial Intelligence, and Modeling & Simulation.

Anjanette R. Lasala is currently taking up Bachelor of Science in Computer Science at the Polytechnic University of the Philippines, Mabini Campus. She is knowledgeable in C, C#, Java, HTML, CSS, PHP, SQL and MATLAB. She is interested on Web Development and Database Management.