**Chapter I – The Problem: Rationale and Background**

**Introduction**

Inappropriate Content has been defined in the Children's Internet Protection Act as visual depictions that are obscene, child pornography, or material that is "harmful to minors." Categories under this topic include pornography, hate groups, violence, illegal activity, extremist groups, drugs and online advertising [Inappropriate Content, n.d.].

In this study oﬀensive language is deﬁned as the propagation of oﬀensive messages or remarks that in some circumstances are inappropriate, exhibit a lack of respect towards certain groups of people or are just rude in general [Vandersmissen, 2012].

The Internet has proven a useful tool for pedophiles and sexual predators as they distribute child pornography, engage in sexually explicit conversations with children, and seek victims in chat rooms. The more pornography these individuals access, the higher the risk of their acting out what they see, including sexual assault, rape, and child molestation [Hughes, 1998].

Classifying Inappropriate Content in English Text is important because many people who uses the internet can read inappropriate text content. Those inappropriate contents are very harmful to the readers, it can bully, harassment or etc. [Inappropriate Content, n.d.].

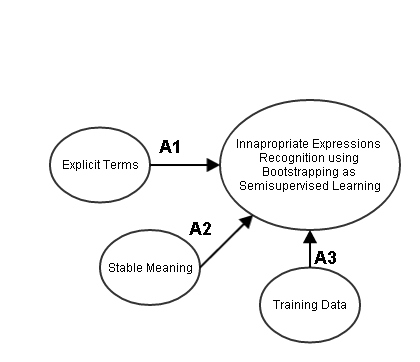
The algorithm to be use in this study is Bootstrapping as Semi-Supervised Learning. For Natural Language, it shows incorporating features derived from unlabeled data into a supervised model can provide substantial improvements, both in terms of reducing the error and the amount of labeled data required. Its results show that using word clusters and a new type of unlabeled data feature, mutual information statistics, can both boost performance [Liang, 2005].

**Statement of the Problem**

The study aims to design, develop and evaluate the system which will help to recognize Inappropriate Expressions from a document. The people who are in knowledgeable in English language are the respondents in this study. In addition to this, the researchers aim to seek answer to the following specific problems:

1. What is the performance analysis of Sentiment Analysis based on the following:
   1. Accuracy
      1. Precision
      2. Recall
      3. Specificity
      4. F-Score

**Conceptual Framework**

****

**A1:** The More Explicit Terms there are in the document, the higher the accuracy

**A2:** The More Stable the Meaning in the Sentences in the document, the higher the accuracy

**A3:** The More Training Data there is in the System, The higher the Accuracy

**Significance of the Study**

The reason why this study is significant can be explained from two aspects:

First, it is significant to students who have a whole society around technology because most of them routinely use chat and email to communicate with each other. The system will prevent receiving possible inappropriate expressions through the internet.

Second, it is also significant to the parents who are worrying so much about what their children do on the internet because some articles found on google are not appropriate for their child and contains inappropriate expressions like offensive, drug related and sexually explicit words.

**Scope and Limitations of the Study**

The study will focus on recognizing inappropriate expressions such as offensive and sexually explicit expressions on a document. This system will also check for the context in the sentence and the language that this system will only consider is in English language.

**Definition of Terms**

Bootstrapping – refers to the process of loading the basic software into the memory of a computer after power on or general reset; especially the operating system which will then take care of loading other software as needed.

Inappropriate Expression – is something that is not within the bounds of what is considered appropriate or socially acceptable.

Natural Language Processing – is a field of computer science, artificial intelligence and computational linguistics concerned with the interactions between computers and human languages.

Offensive Language – the term that is applied to hurtful, derogatory or obscene comments made by one person to another person.

Semisupervised Learning – is a class of supervised learning tasks and techniques that also make use of unlabeled data for training – typically a small amount of labeled data with a large amount of unlabeled data.

Sentiment Analysis – refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials.

Sexually explicit – presents sexual content without deliberately obscuring or censoring it.

**Chapter 2: Review of Related Literature**

Profanity is an offensive word or offensive language [Merriam-Webster Online Dictionary, 2014]. It is also called bad language, strong language, coarse language, foul language, bad words, vulgar language, lewd language, swearing, cursing, cussing, or using expletives. This use is a subset of a language's lexicon that is generally considered to be strongly impolite, rude or offensive. It can show adebasement of someone or something, or show intense emotion. Profanity in this sense takes the form of words or verbal expressions. In its older, more literal sense, the term profanity refers to offensive words, or religious words, used in a way that shows you do not respect God or holy things, or behaviour showing similar disrespect [Longman Dictionary of Contemporary English, 2014].

An Analysis of the Pragmatic Functions of Swearing in Interpersonal Talk has shown the reason why people would choose to swear and the types of pragmatic functions which swearing carries out in everyday conversation. These functions include expressing positive emotions, including showing surprise, promoting in group membership, verbal emphasis to emphasize the speaker’s feeling about something and negative emotions, such as aggression, which threaten a person’s positive and/or negative face [Wang, 2013].

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 [Bi, n.d.].

Lexical Syntactical Feature (LSF) approach to identify offensive contents in social media, and further predict a user’s potentiality to send out offensive contents. Our research has several contributions. First, we practically conceptualize the notion of online offensive contents, and further distinguish the contribution of pejoratives/ profanities and obscenities in determining offensive contents, and introduce hand authoring syntactic rules in identifying name-calling harassment. Second, we improved the traditional machine learning methods by not only using lexical features to detect offensive languages, but also incorporating style features, structure features and context-specific features to better predict a user’s potentiality to send out offensive content in social media. 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. Besides, the LSF tolerates informal and misspelling contents, and it can easily adapt to any formats of English writing styles. We believe that such language processing model will greatly help online offensive language monitoring, and eventually build a safer online environment [Chen et.al., 2012].

Very few other research teams are working on the detection of cyber bullying. A misbehaviour 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% [Yin et. al., 2009].

Filtering Offensive Language in Online Communities using Grammatical Relations has a proposed semantic ﬁltering technique based on the grammatical relations of words in a sentence so that the rest of the ﬁltered sentence is readable and the existence of oﬀensive words in the original sentence is hard to notice. We tested the eﬀectiveness of our approach with a large dataset and the results show that our techniques are very eﬀective and accurate with little process overhead [Xu et. al., 2010].

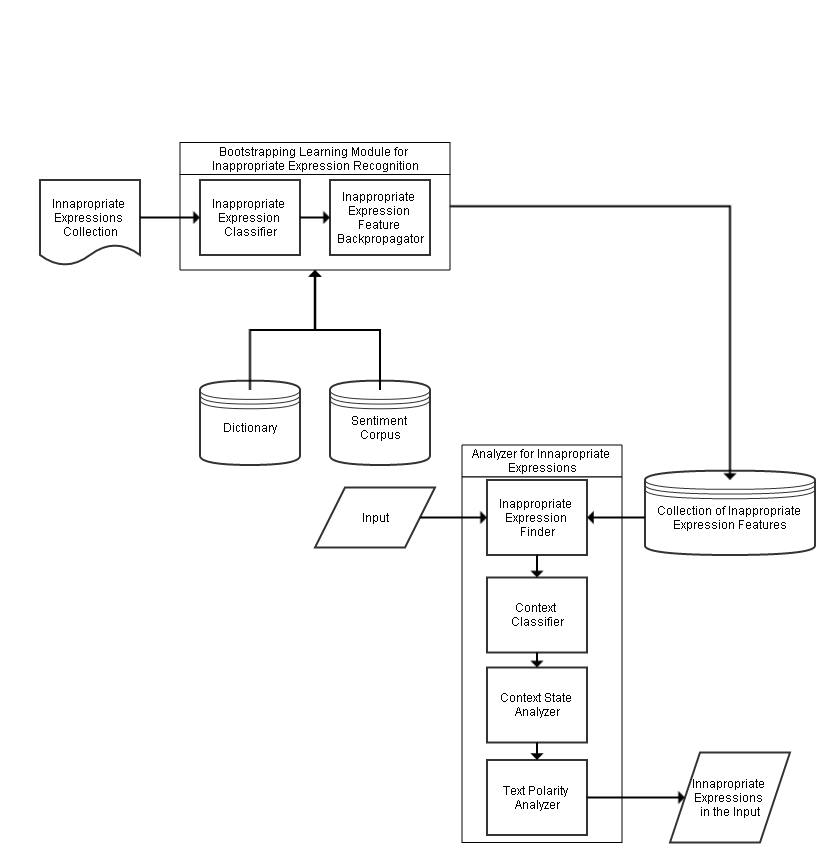
Semi-supervised learning is a class of supervised learning tasks and techniques that also make use of unlabeled data for training - typically a small amount of labeled data with a large amount of unlabeled data. Semi-supervised learning falls between unsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data). Many machine-learning researchers have found that unlabeled data, when used in conjunction with a small amount of labeled data, can produce considerable improvement in learning accuracy [Wikipedia, n.d.].

Semi-Supervised Learning for Natural Language shows incorporating features derived from unlabeled data into a supervised model can provide substantial improvements, both in terms of reducing the error and the amount of labeled data required. Its results show that using word clusters and a new type of unlabeled data feature, mutual information statistics, can both boost performance. In addition, semi-Markov models can also increase performance modestly on the named-entity recognition (NER) task but in some cases hurts performance on the Chinese word segmentation (CWS) task [Liang, 2005].

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 [Li et. al., n.d.].

**Chapter 3 – Research Methodology**

**System Architecture**



# References

[1] Inappropriate Content (<https://sites.google.com/site/nbushra100/home/elizabeth>)

[2] Abney, S. (2008). Semisupervised Learning for Computational Linguistics.

(<https://books.google.com.ph/books?id=VCd67cGB_rAC&pg=PA4&lpg=PA4&dq=problems+in+semi+supervised+learning&source=bl&ots=Up8BGfWB62&sig=huquA4-qDJXSUpNrnkdH1T_UnL4&hl=en&sa=X&ved=0CGcQ6AEwCWoVChMIxOK6qIaSxwIVUFKOCh3oogiW#v=onepage&q=problems%20in%20semi%20supervised%20learning&f=false>)

[3] Blum, A., &Balcan, M. (n.d.) Open Problems in Efficient Semi-supervised PAC Learning.

(http://www.cs.cmu.edu/~ninamf/papers/openproblems.pdf)

[4] Vandersmissen, B. (2012). Automated Detection of oﬀensive language behavior on social networking sites. (<http://lib.ugent.be/fulltxt/RUG01/001/887/239/RUG01-001887239_2012_0001_AC.pdf>)

[5] Hughes, D.R. (1998). How Pornography Harms Children.

(http://­­www.protectkids.com/effects/harms.htm)

[6] "Definition of Profanity", Merriam-Webster Online Dictionary, retrieved on 2014-08-31

[7] "Definition of profanity". Longman Dictionary of Contemporary English – online. (Accessed on 11 September 2014)

[8] Wang N. (2013). An Analysis of the Pragmatic Functions of Swearing in Interpersonal Talk. (https://www.griffith.edu.au/\_\_data/assets/pdf\_file/0007/589453/Na-Wang.pdf)

[9] 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>)

[10]Chen, et. al. (2012). Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. (http://www.cse.psu.edu/~sxz16/papers/SocialCom2012.pdf)

[11] Dawei Yin, ZhenzhenXue, Liangjie Hong (2009).*Detection of Harassment on Web 2.0.* Madrid, Spain.(<http://www.cse.lehigh.edu/~brian/pubs/2009/CAW2/harassment.pdf>) (accessed on April 21, 2009)

[12] Xu, Z., &Zhu, S. (2010) Filtering Offensive Language in Online Communities using Grammatical

Relations. (<http://ceas.cc/2010/papers/Paper%2010.pdf>) (accessed on July 14, 2010)

[13] Wikipedia (n.d.). Semi-Supervised Learning. (

[https://en.wikipedia.org/wiki/Semi](https://en.wikipedia.org/wiki/Semisupervised_learning)

[supervised\_learning](https://en.wikipedia.org/wiki/Semisupervised_learning))

[14] Liang, P. (2005). Semi-Supervised Learning for Natural Langauge.

(<http://cs.stanford.edu/~pliang/papers/meng-thesis.pdf>) (accessed on May 2005)

[15] Li et. al. (n.d.). Semi-Supervised Learning for Imbalanced Sentiment Classification.

(<https://www.aaai.org/ocs/index.php/IJCAI/IJCAI11/paper/download/3131/3744>.)