# CHAPTER 1: THE PROBLEM AND ITS BACKGROUND

## **1.1 Introduction­­**

Inappropriate Expressions is one of the p­­roblems in a behavioral sense. 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 solution to be used in this study is a machine learning methodology, which is bootstrapping, which is designed to model the 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.

## **1.2 Background of the Study**

In many ways the Internet is like a gigantic library; both have content to teach and entertain. And similar to the content in a library, not all Internet content is appropriate for children. Libraries create children’s and young adults’ sections in order to help youths (and their parents) identify which materials are appropriate for them. On the Internet, however, all of the content may be equally accessible; websites about ponies and websites featuring pornography are both a click away. 87% of children go online at home and it is possible that they can read some inappropriate expressions on the internet. This inappropriate expressions could be expressions containing swearing, unmoderated chatrooms where there’s no one supervising the conversation and barring unsuitable comments and sexual explicit comments [Inappropriate Content, 2015].

This inappropriate expressions becomes very interesting in the field of NLP Community. Fortunately, some researches like CAW 2.0 build a system that detects cyber bullying on the internet. 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% [Hong et. al., 2009] . And another research that was built was profanity related offensive content in twitter. The researchers 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. 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 [Fan et.al, n.d.]. Therefore, one of the problems that need to be solve in this study is the accuracy of the system because some of the researches related to the study has a low rate of accuracy.

In Sentiment classification, it aims to predict the sentiment polarity of a text [Lee et al., 2002] and plays a critical role in many Natural Language Processing (NLP) applications [Cheng et al., 2005; Cardie et al., 2005; Cui et al., 2006; Balahur et al., 2009]. Although supervised learning methods have been widely employed and proven effective in sentiment classification in the literature [Lee et al., 2002], they normally depend on a large amount of labeled data, which usually involves high cost in labor and time. To overcome this problem, various semi-supervised learning methods are proposed to effectively utilize a small scale of labeled data along with a larger amount of unlabeled data.

One specific open problems in semi-supervised learning is the co-training with linear separators which is known that the consistency problem is NP-hard. Even if one cannot solve the problem efficiently in general, a natural question is whether one can at least weaken the independence given the label assumption in a non-trivial way and still get an efficient algorithm for this class [Balcan and Blum, n.d.].

However, all the existing semi-supervised learning methods assume the balance between negative and positive samples in both the labeled and unlabeled data, and none of them consider a more common case where the class distribution is imbalanced, i.e., the number of positive samples is quite different from that of negative samples in both the labeled and unlabeled data. For clarity, the class with more samples is referred as the majority class (MA) and the other class with fewer samples is referred as the minority class (MI). In fact, semi-supervised learning on imbalanced classification is rather challenging: at least, there exist two basic issues to be solved. On the one hand, imbalanced classification requires a specifically-designed classification algorithm. Trained on the imbalanced labeled data, most classification algorithms tend to predict test samples as the majority class and may ignore the minority class. Although many methods, such as re-sampling [Bowyer et al., 2002], one-class classification [Duin and Juszczak, 2003], and cost-sensitive learning [Liu and Zhou, 2006], have been proposed to solve this issue, it is still unclear as to which method is more suitable to handle the imbalanced problem in sentiment classification and whether the method is extendable to semi-supervised learning.

The solution to be used in this study is a machine learning methodology, which is bootstrapping, which is designed to model the 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.

## **1.3 Theoretical Framework**



The theory, which is Neuro-Psycho-Social Theory of Speech. There are “Rules” that governs the human’s expression of Inappropriateness, Offensiveness, and Humor. The catalysts of these expressions are the Neurological Factors, Psychological Factors, and Social-Cultural Factors. [Jay, 2009]

## **1.4 Conceptual Framework**

### **1.4.1 Conceptual Framework of the System**

****

The Concept is to create a model in which there will be a. machine that learns inappropriate expressions and recognize it. The process of training will be by the utilization of a sentiment corpus to represent the catalysts of the Inappropriate expression, which contains the sentiment polarity values, which will affect the threshold values for the inappropriateness and the features that will be collected in the inappropriate expressions feature collection, in which leads to affect the performance of the recognition of inappropriate expressions during the simulation.

### **1.4.2 Conceptual Framework of the Study**

****

­­­­­­­ The Concept is to study a model in which there will be a. machine that learns inappropriate expressions and recognize it. There will be different training sets and different sample documents. Then, the results will be experimented via Experiment Paper. The output will be the evaluations and recommendation of the system and the approach to solve the problem about recognizing inappropriate expressions.

## **1.5 Statement of the Problem**

The study aims to design, create and evaluate the Model that recognizes Inappropriate Expressions from a document. In addition to this, the researchers aim to seek answer to this problem:

What is the performance analysis of the model in terms of:

1. Recognition of Inappropriate expressions. (Accuracy of the Model)

2. Recognition of Appropriate expressions. (Specificity of the Model)

## **1.6 Scope and Limitations**

### **1.6.1 Scope and Limitation of the System**

The system will be deployed as a Java Application. The system will be dependent of the tools such as WordNet, SentiWordNet, Stanford CoreNLP, and MIT JWI Stemmer. The system that will be developed will be dependent on the Latest Java Virtual Machine and Runtime Environment and a Python Interpreter. There will be also a component for the scraping of Urban Dictionary definitions with the use of Python with the requests module and beautifulSoup.

The system’s input consists of a textual input. Underscores and slashes symbols are to be avoided in the input to avoid conflicts with the Stanford CoreNLP.

The algorithms to be used is Bootstrapping for the machine learning, Naïve Bayes Text Classification for the analysis of the inappropriate expressions and for the feature extraction in the inappropriate expressions learning process.

### **1.6.2 Scope and Limitations of the Study**

The time frame for the development of the system will compose of an estimated time of 3 months of prototyping. The study will be evaluated by an expert, which is an English teacher. The inputs will be documents and text file.

## **1.7 Significance of the Study**

Inappropriate Expressions Recognition using Bootstrapping as Semi-supervised Learning focuses on recognizing inappropriate expressions on a document. The system will benefit the following:

**Students** – these are people 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.

**Parents** – these are the people 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 and sexually explicit words.

**Editors-** these are the people who are worrying so much about the works that will be published. With this model, there will be recognition of inappropriate expressions in the document that may help the editorial judge the work’s level of inappropriateness.

**Moderators-** these are the people who manage the behavior of users in an online community. With this model, the moderators can have ease in reading the activities that has to deal with inappropriate expressions.

**Natural Language Processing Researchers (NLP)** – These are the experts in Natural Language processing. The system will provide avenues for further improvement of the said topic.

**Future Researches** – These are the people who will conduct future research to improve the existing studies about Sentiment Analysis and Inappropriate Expressions. They can add additional functions to the system.

## **1.8 Definition of Terms**

Bootstrapping Algorithm– is a method for deriving robust estimates of standard errors and confidence intervals for estimates such as the mean, median, proportion, odds ratio, correlation coefficient or regression coefficient.

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

Naive Bayes - methods are a set of supervised learning algorithms based on applying Bayes’ theorem with the “naive” assumption of independence between every pair of features.

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.

Part-of-Speech Tagger -  is a piece of software that reads text in some language and assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc., although generally computational applications use more fine-grained POS tags like 'noun-plural'.

Named Entity Recognizer - is a piece of software that reads text in some language and assigns a classification of proper nouns such as person, location, organization, etc.

N-Gram - is a contiguous sequence of words from a given sequence of text.

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.

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

SentiWordNet - is a lexical resource for opinion mining. SentiWordNet assigns to each synset of WordNet three sentiment scores: positivity, negativity, objectivity.

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

Urban Dictionary – is comprised of submissions from everyday people and regulated by volunteer editors, who are given an extremely small set of rules to maintain quality control.

WordNet - is a lexical database for the English language. It groups English words into sets of synonyms called synsets, provides short definitions and usage examples, and records a number of relations among these synonym sets or their members.

# CHAPTER 2: REVIEW OF RELATED LITERATURE AND STUDIES

This chapter includes review of related literature and studies. It also includes the synthesis of the study. This chapter provided relevant information and researchers regarding the study that gave additional knowledge to the researchers.

## **2.1 Review of Related Literature**

**2.1.1 Why we swear or curse?**

Swear words (or taboo words, as he calls them) can include sexual references (fuck), those that are profane or blasphemous (goddamn), scatological or disgusting objects (shit), animal names (pig, ass), ethnic/racial/gender slurs (fag), ancestral allusions (bastard), substandard vulgar terms and offensive slang. Taboo words can be mildly offensive to extremely offensive, and people will often use a more mild euphemism to replace a swear word when in mixed (or unknown) company [Jay, 2009].

Taboo words can be used for a variety of reasons, including to achieve a specific reaction from others. Swearing injects a direct, succinct emotional component into the discussion, usually in order to express frustration, anger or surprise (up to two-thirds of our swearing is for just such expressions). These insulting swears can be name calling or wishing someone harm, so it’s no wonder they are often a defining feature of hate speech, verbal abuse, sexual harassment and obscene phone calls [Grohol, 2009].

Cursing appears as soon as children hear curse words, as early as one year of age [Jay, 2009]. Children's cursing emerges in a predictable fashion. Early cursing and name calling are based on references to scatology and perceived differences about others (e.g., not-eater, four-eyes).In adolescence, cursing becomes more abstract and socially based. Gender differences in cursing emerge as soon as children attend school: Boys curse more and use more words and use more offensive words than do girls. Cursing reaches a peak in adolescence but continues into old age, persisting through senile decline and dementia.

Children learn that curse words are associated with emotion states through classical conditioning, the repeated pairing of words (e.g., damn!) with emotional events. Curse words effectively replace infantile expressions of anger such as biting and screaming [Goodenough, 1931]. Children associate curse words with all emotion states (e.g., joy, surprise, fear); they learn to express emotions through words, and they learn to perceive others' emotional states through the emotional speech they observe. Children learn that curse words intensify emotions in a manner that noncurse words cannot achieve.

Language learning and cursing depend on one's social, emotional, and cognitive reasoning abilities. As children become more cognitive1y sophisticated, their emotional language, name calling, and sexual references shift to match their higher mental functioning. Many uses of curse words occur at an automatic or reflexive level in the form of response cries and epithets. Eventually, the semantic and syntactic rules for cursing are acquired, allowing children to use curse words appropriately in propositional statements [Jay, 2009]. These propositional statements are primarily used to express emotions (connotation), but curse words also function to make references about the world (denotation).

Cursing is never chaotic, meaningless, or random behavior – cursing is seen as purposeful and rule-governed. The rule of NPS Theory is to generate a likelihood “rules” that underlie concepts of appropriateness, offensiveness and humor. Native speakers acquire cursing rules as they learn language. Discovering and testing these cursing rules is meant to give the theory predictive power. The more accurately the NPS Theory can predict acts of cursing, the more valid is our understanding of cursing [Jay, 2009].

**2.1.2 Sexual Identity and Sexual Terminology**

Human sexuality is a critical aspect of emotional language in general and of cursing in particular because sexuality is one of the most tabooed aspects of human existence. The language of sexuality is intimately connected to one's emotional life, one's sexual orientation, and one's cursing habits or style. Human sexuality becomes represented in two ways: The sexual body is represented as a materiality, and a set of sexual ideas or sexual language is developed about that materiality [Jay, 2009].

Children learn sexual terminology through interactions with peers and adults. Parents express their sexual values, fears, and anxieties to children when they inhibit or punish sexual references. Punishment and avoidance of sex terms teach the child that sexual words are powerful and that sexuality itself is powerful. Parents with high sexual anxiety are likely to transfer their anxiety to their children, who learn that both sex talk and sex are to be avoided. This learning takes place through the repression of sex talk itself through a course of negations and omissions. Through the acquisition of sexual terminology and the conditioned fears and pleasures regarding sexuality, the child develops a level of comfort with sexuality. This sexual identity will influence how a speaker uses words for sex acts, body parts, and gender related insults with other people [Jay, 2009].

As children develop linguistically and sexually, their conversations about sexuality become more highly dependent on who is listening. Both adolescent and adult sexual conversations clearly depend on intimacy, sexual identity, and formality [Wells, 1990]. Most adults can talk about sexuality with lovers or with others who share similar sexual preferences. But almost all young adults have trouble talking about sex with their parents (those who avoided sex talk in the first place) and in mixed gender crowds.

Most people talk about sex by using vulgar terms and sexual slang [Jay, 2009]. Clinical terms are reserved for polite situations. Some sex acts are so taboo (e.g., oral sex) that no acceptable term can be used in polite company. Euphemisms and circumlocutions are commonly used in order to talk about sex and taboo topics. In fact, married couples, cohabiting couples, and sexually active partners create personal idioms and idiosyncratic terminology to use in intimate situations [Cornog, 1986]. We see in the following chapters in Part ill that the use of sexual language is very important to speakers, revealing their personality traits, attitudes about sexuality, and parental influences. Both the physical acts of sex and one's sexual identity are expressed and experienced through language choices.

**2.1.3 Inappropriate Expression**

Inappropriate Expression is something that is not within the bounds of what is considered appropriate or socially acceptable [YourDictionary, n.d.]. Inappropriate expression includes information that upset us or information that might lead or tempt us into unlawful or dangerous behavior. This could be content containing swearing, unmoderated chatrooms where there’s no one supervising the conversation and barring unsuitable comments and sexual explicit comments [Inappropriate Content, [n.d.]. Profanity is an offensive word or inappropriate 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 a debasement 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 behavior showing similar disrespect [Longman Dictionary of Contemporary English, 2014].

Words become inappropriate in few overlapping ways.

**2.1.4 IBM’s Super Computer learns to Swear**

The first cognitive system was Watson, which debuted in a televised Jeopardy! Challenge where it bested the show’s two greatest champions. The challenge for Watson was to answer questions posed in every nuance of natural language, such as puns, synonyms and homonyms, slang, and jargon [IBM Research, n.d.].

In 2011 Watson, the IBM super computer best known for its run as a Jeopardy contestant, gained a new tool in its language arsenal: swearing [How-to-Geek, n.d.]. Researchers in charge of expanding Watson’s vocabulary and ability to use language in a more nuanced and natural fashion thought it would be helpful to teach Watson slang and colloquial sayings. Essentially, they wanted to give Watson the ability to speak more like we speak to each other and less like a super computer carefully selecting an answer [How-to-Geek, n.d.].

To this end the researchers unleashed Watson on Urban Dictionary, the massive 7-million-entry Internet dictionary of slang words and phrases. On one hand the experiment was a huge success, crunching through the Urban Dictionary database radically expanded Watson’s word selection and the nuance of its language use. On the other hand, Watson proved to be terrible at distinguishing when its language was and was not appropriate–shortly after it acquired its new vocabularly from Urban Dictionary it reportedly responded to a researcher’s inquiry by reporting it was “bullshit” [How-to-Geek, n.d.].

In response to the changes in Watson’s vocabulary researchers terminated the Urban Dictionary experiment and set up filters to help Watson refrain from swearing in the future [How-to-Geek, n.d.].

**2.1.5 Word Filter**

A *swear filter*, also known as a *profanity filter* or *language filter* is a software subsystem which modifies text to remove words deemed offensive by the administrator or community of an online forum. Swear filters are common in custom-programmed chat rooms and online video games, primarily MMORPGs. This is not to be confused with content filtering, which is usually built into internet browsing programs by third-party developers to filter or block specific websites or types of websites. Swear filters are usually created or implemented by the developers of the Internet service [FileSharingTalk, 2006].

A common quirk with wordfilters, often considered either comical or annoying by users, is that they often affect words that are not intended to be filtered. This is a typical problem when short words are filtered. For example, if the word "ass" is filtered, so are "assist", "classic", "assassin", and other words which contain the sequence. For example, one may see, "Do you need \*\*\*istance for playing cl\*\*\*ical music?" Multiple words may be filtered if whitespace is ignored, resulting in "as suspected" becoming "\*\*\*\*uspected". Prohibiting a phrase such as "hard on" will result in filtering innocuous statements such as "That was a hard one!" and "Sorry I was hard on you." [Sheerin, 2010].

Some words that have been filtered accidentally can become replacements for profane words. One example of this is found on the [Myst](https://en.wikipedia.org/wiki/Myst) forum Mystcommunity. There, the word 'manuscript' was accidentally censored for containing the word 'anus', which resulted in 'm\*\*\*\*cript'. The word was adopted as a replacement swear and carried over when the forum moved, and many substitutes, such as " 'scripting ", are used (though mostly by the older community members) [Sheerin, 2010].

Place names may be filtered out unintentionally due to containing portions of swear words. In the early years of the internet, the British place name [Penistone](https://en.wikipedia.org/wiki/Penistone) was often filtered out from spam and swear filters [Sheerin, 2010].

**2.1.6 Bootstrapping Algorithm**

The Bootstrap algorithm works by drawing many independent bootstrap samples, evaluating the corresponding bootstrap replications, and estimating the standard error of by the empirical standard error, denoted by B, where B is the number of bootstrap samples used [Efron and Tibshirani, 1993].

With the bootstrap method, the basic sample is treated as the population and a Monte Carlo-style procedure is conducted on it. This is done by randomly drawing a large number of ‘resamples’ of size n from this original sample (of size n either) with replacement. So, although each resample will have the same number of elements as the original sample, it could include some of the original data points more than once, and some not included. Therefore, each of these resamples will randomly depart from the original sample. And because the elements in these resamples vary slightly, the statistic G\*, calculated from one of these resample will take on slightly different values. The central assertion of the bootstrap method is that the relative frequency distribution of these G\*’s is an estimate of the sampling distribution of G [The Original Bootstrap Method, n.d.].

**2.1.7 Semi-supervised Learning**

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].

**2.1.8 Naïve Bayes Classifier**

Naive Bayes has been studied extensively since the 1950s. It was introduced under a different name into the text retrieval community in the early 1960s, [Norvig and Russell et.al., 2003] and remains a popular (baseline) method for text categorization, the problem of judging documents as belonging to one category or the other (such as spam or legitimate, sports or politics, etc.) with word frequencies as the features. With appropriate preprocessing, it is competitive in this domain with more advanced methods including support vector machines [Karger et.al. 2003]. It also finds application in automatic medical diagnosis [Rish, 2001].

Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression [Norvig and Russell, 2003], which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.

## **2.2 Review of Related Studies**

**2.2.1 Sentiment Analysis in Classifying Offensive Language**

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.].

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 [Bellmore et.al, 2012].

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

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 smodel, a support vector machine learner could only produce a recall level of 61.9% [Hong et. al., 2009].

In the research Filtering Offensive Language in Online Communities using Grammatical Relations, mainly tackles the problem about how the offensive language can be removed in a user message. They analyze the offensive language in text messages posted in online communities, and propose a new automatic sentence-level filtering approach that is able to semantically remove the offensive language utilizing the grammatical relations among words. Their solution includes 3 steps. First, they analyzed the characteristics of offensive text content in user messages. Then, they introduced their filtering philosophy according to the summarized characteristics. Finally, they show how this philosophy is transformed into heuristic rules applicable in the filtering process. The overview idea of their filtering approach is shown in the presented Algorithm 1 in which the functions POS tagging ad TD generator generate Part-of-Speech tags and typed dependency relations, respectively. They also use existing NLP (Natural Language Processing) tools to implement these two functions. They also focused in the design of two other functions CreateRelTree and EstimateRelTree. In their research assume that the filtering is based on a comprehensive offensive lexicon containing all offensive words. Words do not appear in the lexicon are considered inoffensive. Experiments their dataset, comments from Youtube, show over 90% agreement in filtered results between the proposed approach and manual filtering approach [Xu and Sencun, 2010].

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

**2.2.2 Semi-supervised Sentiment Classification**

Generally, sentiment classification methods can be categorized into three types: unsupervised [Turney, 2002], supervised [Pang et al., 2002], and semi-supervised [Melville and Sindhwani, 2008]. 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.

One kind of semi-supervised methods for sentiment classification is to utilize prior lexical knowledge in conjunction with the labeled and unlabeled data. For example, Document-Word Co-regularization for Semi-supervised Sentiment Analysis analyzed the sentiment of documents and words based on a bipartite graph representation of the labeled and unlabeled data while Li et al. [2009] employed some simple update rules to make use of tri-factorization of the term-document matrix. It is rather common that such methods require a high-quality lexicon with the polarity of words properly defined [Melville and Sindhwani, 2008].

Another kind of semi-supervised methods for sentiment classification is to employ some bootstrap techniques, such as self-training [Yarowsky, 1995] and co-training [Blum and Mitchell, 1998]. Among them, co-training has been proven more effective than self-training [Wan, 2009; Huang et.al, 2010]. The key issue of applying co-training is to find a suitable set of different views. For instance, Co-Training for Cross-Lingual Sentiment Classification regarded two different languages (i.e., English and Chinese) as two views [Wan, 2009] while Employing Personal/Impersonal Views in Supervised and Semi-supervised Sentiment Classification considered personal and impersonal texts as two views [Huang et.al, 2010]. This paper employs the co-training technique and generates different views from random feature subspaces. Among others, Mine the Easy and Classify the Hard: Experiments with Automatic Sentiment Classification integrated various methods, such as spectral clustering, active learning, transductive learning, and ensemble learning, in semi-supervised sentiment classification [Dasgupta and Ng, 2009]. To our best knowledge, no existing semi-supervised methods consider the class imbalance problem in sentiment classification.

Semi-Supervised Learning for Semantic Relation Classification using Stratified Sampling Strategy explores several key issues in semi-supervised learning based on bootstrapping for semantic relation classification. The application of stratified sampling originated from statistics theory to the selection of the initial seed set contributes most to the performance improvement in the bootstrapping procedure. In addition, the more strata the training data is divided into, the better performance will be achieved. However, the augmentation of the labeled data using the stratified strategy fails to function effectively largely due to the unbalanced distribution of the confidently classified instances, rather than the stratified sampling strategy itself [Kong et.al, 2009].

Semi-supervised Learning for Relation Extraction integrate the advantages of SVM bootstrapping in learning critical instances and label propagation in capturing the manifold structure in both the labeled and unlabeled data, by fir­st bootstrapping a moderate number of weighted support vectors through a co-training procedure from all the available data, and then applying label propagation algorithm via the bootstrapped support vectors [Li et.al, 2008].

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 [Balcan and Blum, n.d.].

**2.3 Synthesis of the Study**

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.

# CHAPTER 3: RESEARCH METHODOLOGY

## **3.1 Research Method Used**

The experimental research method involves manipulating one variable to determine if changes in one variable causes changes in another variable. This method relies on controlled methods, random assignment and the manipulation of variables to test a hypothesis.

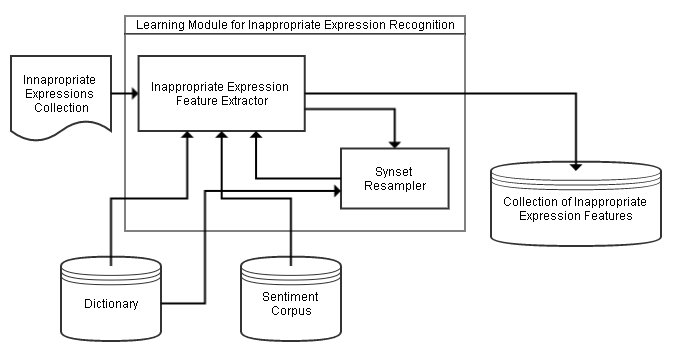
The system will not require respondents since the system uses Experimental Research. Instead, the researchers will assess the performance of the system given implementation of algorithms and techniques. The system will be tested by inputting a document containing inappropriate expressions and calculates its performance in terms of scores in accuracy.

## **3.2 Research Paradigm**

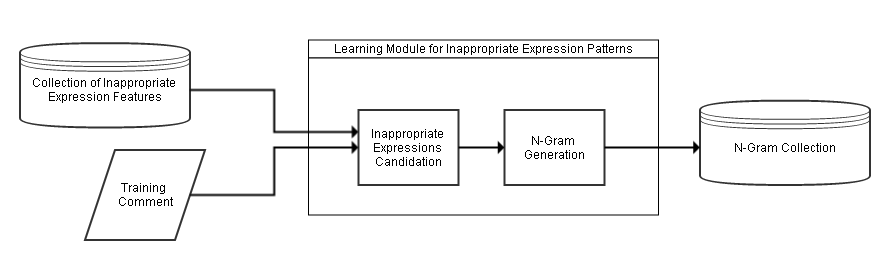
The researchers would implement positivist way of approach. This is because Inappropriate Expressions Recognition requires lot of testing and observational analysis to ensure accurate and better result compare to the other studies. Series of testing and analysis of the system will be done to get the result with the highest level of accuracy.

In this way, the researchers will have the series of data that will act as empirical evidence and later will be used to analyze the system accuracy and performance level.

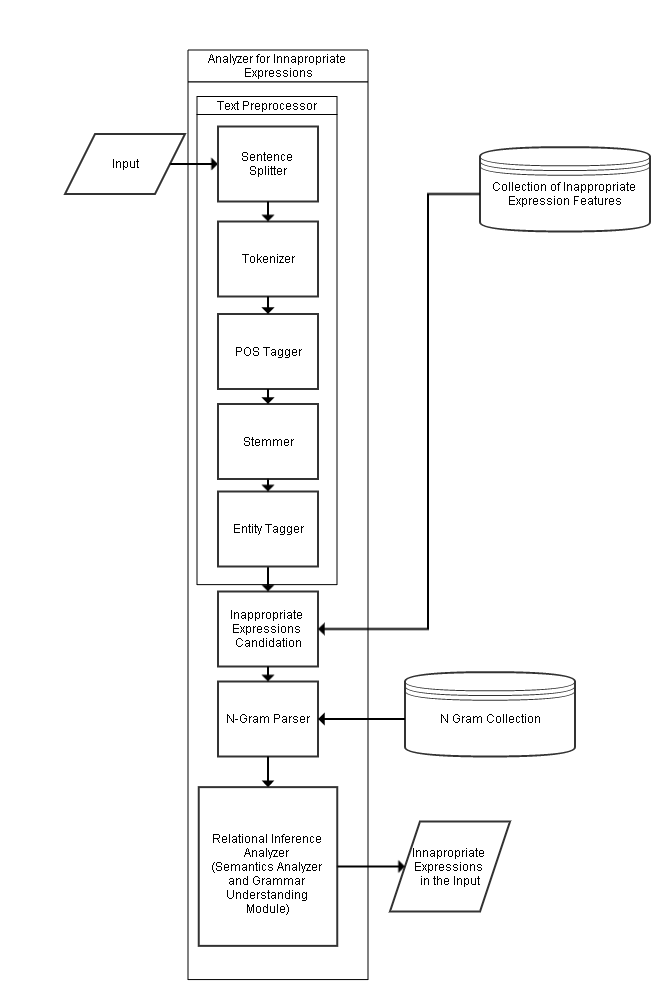
## **3.3 System Architecture**

**­**

In the Training phase there is a 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 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. 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.



There is another 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. 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.



In 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. 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. 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 relations to the other existing words in the same input.

## **3.4 Sampling Technique**

The researchers will implement convenience sampling. Convenience sampling is a non-probability sampling technique where subjects are selected because of their convenient accessibility and proximity to the researcher.

The subjects, which is inappropriate expressions from 9gag comments, youtube.com comments, reddit textual posts and 4chan textual posts, are selected just because they are easiest to recruit for the study and the researcher did not consider selecting subjects that are representative of the entire population.

## **3.5 Description of Subjects**

The subjects which are used in testing the model are 9gag and youtube comments. The reason why the researchers chose those websites is because some of the users engage in that media is using English Language as a medium of communication. Another reason is because of publicity of the content of these sites. The researchers also observed that most of the users there are using inappropriate expressions in expressing their feelings, reactions and opinions.

## **3.6 Instrumentation**

Instrumentation refers to the tools or means by which researchers attempt to measure variables or items of interest in the data collection process. The system will be deployed as a Java Application with some Python scripts embedded in the system and the tools that will be use in developing the system are WordNet, SentiWordNet, Stanford CoreNLP, MIT JWI Stemmer, requests module, and beautifulSoup. This tools will serves as the dictionary and will be use in the pre-processing phase of the system. The system that will be developed will be dependent on the Latest Java Virtual Machine and Runtime Environment, and Python Interpreter.

The study will also utilizes experiments to test its effectiveness on recognition of inappropriate expressions. So as the researchers will be using experiment paper to identify the results of the tests conducted.

## **3.7 Data Gathering Procedure**

The data gathered by the researchers will come from the results in the experiments performed. The experimentation of the model will be done by testing the performance in the recognition of Inappropriate Expressions and Appropriate Expressions.

There are two experiments performed in each leaves:

1) Testing for the Accuracy – This experiment will be done by providing documents with inappropriate expressions in the model and testing it if it recognizes them.

2) Testing for the Specificity - This experiment will be done by providing documents without any inappropriate expressions or inappropriate expressions without inappropriate sense in the model and testing it if it avoids them.

## **3.8 Statistical Treatment**

Statistical treatment consists of formulas that were used to answer what the problem states. The formulas used are as follows:

1. Specificity

The performance of the recognition will bemeasured through the use of Specificity. Specificity is the rate of the results without the condition, which has a negative test result. NLP studies uses specificity to eliminate biased results in the system.

Where:

TN (True Negative) – the system correctly indicated that the input is Appropriate

FP (False Positive) – System determined the input is Inappropriate present, the expected output is Appropriate

1. Harmonic Mean or F-Measure

The performance of the Recognition will be measured through the use of the Harmonic Mean, or f-measure. The f-measure is the weighted average of the values of the Precision and Recall. By multiplying the values by 2 and dividing it by the sum of the Precision and Recall, we can get the harmonic mean of the system. A high F1 score will imply a good performance of the system. The formula for the Harmonic mean is as follows:

Where:

TP (True Positive) = System correctly determined the input is Inappropriate

FP (False Positive) – System determined the input is Inappropriate present, the expected output is not

FN (False Negative) – System indicated that the input is Appropriate, the expected output is it is inappropriate

­­ F = F-measure

P = Precision – Percentage of identified expressions that are inappropriate.

R = Recall – Percentage of inappropriate expressions correctly identified.

# BIBLIOGRAPHY

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

Balahur, M., Lloret, E., Montoyo, A. and Palomar M. Towards Building a Competitive Opinion Summarization System. *In Proceedings of NAACL-09 Student Research Workshop and Doctoral Consortium* (2009), pp. 72-77.

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

Blum, A. and Mitchell, T. Combining Labeled and Unlabeled Data with Co-Training. *In Proceedings of the Eleventh Annual Conference on Computational Learning Theory* (1998), pp. 92-100.

Bowyer, K., Chawla, N., Hall, L. and Kegelmeyer, W. SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research*, 16 (2002), pp. 321-357.

Cardie, C., Wiebe, J. and Wilson T. Anotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation*, 39, 2/3 (2005).

*Chapter 4: The Original Bootstrap Method*. n.d, from VirginiaTech: <http://scholar.lib.vt.edu/theses/available/etd-6169714555/unrestricted/Ch4.pdf>

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

Cheng, J., Hu, M. and Liu, B. Opinion Observer: Analyzing and Comparing Opinions on the Web. *In Proceedings of WWW-05* (2005), pp.342-351.

Cornog, M. Naming Sexual Body Parts: Preliminary Patterns and Implications. *The Journal*

*of Sex Research*, 22, 3 (August 1986), pp. 393-398.

Cui, H., Datar, M. and Mittal, V. Comparative Experiments in Sentiment Classification for Online Product Reviews. *In Proceedings of AAAI-06* (2006), pp. 1265-1270.

Dasgupta, S. and Ng, V. Mine the Easy and Classify the Hard: Experiments with Automatic Sentiment Classification. *In Proceedings of ACL-IJCNLP-09* (2009), pp. 701-709.

"Definition of Profanity", Retrieved from on August 08, 2014 from Merriam-Webster Online Dictionary: <http://www.merriam-webster.com/dictionary/profanity>

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>

Duin R. and Juszczak, P. Uncertainty Sampling Methods for One-Class Classifiers. *In Proceedings of ICML-03, Workshop on Learning with Imbalanced Data Sets II* (2003), pp. 81-88.

Efron, B. and Tibshirani, R.J. *Chapter 8: The Bootstrap*. 1993.

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

FileSharingTalk. Retrieved October 1, 2006 from *When the \*\*\*\* did we get a Word Filter?*: http://filesharingtalk.com/threads/88125-When-the-\*\*\*\*-did-we-get-a-wordfilter

Goodenough, F.L. Anger in Young Children. *Wesport, CT: Greenwood Press* (1931).

Grohol, J. 2009. PsychCentral. Retrieved March 30, 2009 from *Why Do We Swear?*: http://psychcentral.com/blog/archives/2009/03/30/why-do-we-swear/

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

Huang, C., Lee, S., Li, S. and Zhou, G. Employing Personal/Impersonal Views in Supervised and Semi-supervised Sentiment Classification. *In Proceedings of ACL-10* (2010), pp. 414-423.

Inappropriate Content. Retrieved 18 September, 2015 from Ed422 Cybersafety 101: <https://sites.google.com/site/nbushra>

Jay, T. *Why We Curse?* John Benjamins Publishing Company, North Adams, Massachusetts,

2009.

Kong, F., Qian, L., Xhou, G. and Zhu, Q. Semi-Supervised Learning for Semantic Relation Classification using. *In Proceedings of 2009 Conference on Empirical Methods in Natural Language Processing*, 6, 7 (August 2009), pp. 1437-1445.

Lee, L., Pang, B. and Vaithyanathan S. Thumbs Up? Sentiment Classification using

Machine Learning Techniques. *In Proceedings of EMNLP-02* (2002), pp. 79-86.

Li, J.H., Qian, L.H., Zhou, G.D. and Zhu, Q.M. Semi-supervised Learning for Relation Extration. *In Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP-2008)*, 7, 12 (January 2008), pages 32-38.

Liang, P. Semi-Supervised Learning for Natural Language. *In Proceedings of Massachusetts Institute of Technology* (May 2005), pp. 13-73.

Liu, X. and Zhou, Z. Training Cost-Sensitive Neural Networks with Methods Addressing the Class Imbalance Problem. *IEEE Transaction on Knowledge and Data Engineering*, 18 (2006), pp.63-77.

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

Norvig, P. and Russell, S. Artificial Intelligence: A Modern Approach (2nd ed.). *Prentice Hall* (2003).ISBN 978-0137903955.

Rish, I. An Empirical Study of the Naive Bayes Classifier (PDF). *IJCAI Workshop on Empirical Methods in AI* (2001).

Semi-supervised Learning. Retrieved September 18, 2015 from *Wikipedia: The Free Encyclopedia*: <https://en.wikipedia.org/wiki/Semi-supervised_learning>

Sheerin, J. 2010. BBC News. Retrieved March 29, 2010 from *How spam filters dictated Canadian magazine's fate*: http://news.bbc.co.uk/2/hi/technology/8528672.stm

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

Wan, X. Co-taining for Cross-Lingual Sentiment Classification. *In Proceedings of ACL IJCNLP-09* (2009), pp. 235-243.

Watson. from *IBM Research*: http://www.research.ibm.com/cognitivecomputing/watson/

index.shtml#fbid=\_V9cRK7OhZC

Wells, G. Talk about Text: Where Literacy is Learned and Taught. *Curriculum Inquiry*, 20, 4 (1990), pp. 369-405.

Where Did IBM’s Super Computer Watson Learn to Swear?. From *How-to-Geek:*

*http://www.howtogeek.com/trivia/where-did-ibms-super-computer-watson-learn-to-swear/*

Xu, Z. and Sencun Z. Filtering Offensive Language in Online Communities using Grammatical Relations. *In Proceedings of CEAS 2010 - Seventh Annual Collaboration, Electronic Messaging, Anti-abuse and Spam Conference* (July 2010).

Yarowsky, D. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. *In Proceedings of ACL-05* (1995), pp. 189-196.

# APPENDIX A: EXPERIMENT PAPER

1. Experiment Paper I – Determining inappropriate expressions between system and expert.

|  |  |  |
| --- | --- | --- |
| Sentences | Inappropriate | |
| System | Expert |
| 1 |  |  |
| 2 |  |  |
| 3 |  |  |
| 4 |  |  |
| 5 |  |  |
| 6 |  |  |
| 7 |  |  |
| 8 |  |  |
| 9 |  |  |
| 10 |  |  |

1. Experiment Paper II – Input Scoring

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sentence | TP | FP | TN | FN |
| 1 |  |  |  |  |
| 2 |  |  |  |  |
| 3 |  |  |  |  |
| 4 |  |  |  |  |
| 5 |  |  |  |  |
| 6 |  |  |  |  |
| 7 |  |  |  |  |
| 8 |  |  |  |  |
| 9 |  |  |  |  |
| 10 |  |  |  |  |
| Total |  |  |  |  |
| Average |  |  |  |  |

# APPENDIX B – SCREENSHOT

1. Prototype of the System