Vision and Scope Document

for

Automated Detection of Cyberbullying Occurrences in Social Media Posts through Text Classification Using Support Vector Machine (SVM) Algorithm

Version 1.0 approved

Prepared by Samantha Mallari, Faith Ballesteros and Eva Samillano

Asia Pacific College – School of Computing and Information Technologies

October 13, 2016

Table of Contents

Table of Contents ii

Revision History ii

1. Business Requirements 1

1.1. Background 1

1.2. Business Opportunity 1

1.3. Business Objectives and Success Criteria 1

1.4. Customer or Market Needs 1

1.5. Business Risks 1

2. Vision of the Solution 2

2.1. Vision Statement 2

2.2. Major Features 2

2.3. Assumptions and Dependencies 2

3. Scope and Limitations 2

3.1. Scope of Initial Release 2

3.2. Scope of Subsequent Releases 2

3.3. Limitations and Exclusions 3

4. Business Context 3

4.1. Stakeholder Profiles 3

4.2. Project Priorities 4

Revision History

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Date** | **Reason For Changes** | **Version** |
|  |  |  |  |
|  |  |  |  |

# Business Requirements

## Background

Social media is defined as an electronic form of communication wherein people can create, share, and exchange information in the virtual community. It has significantly increased the communication platforms. However, as social media become widespread it has also brought a societal problem that deserves utmost attention and must be addressed immediately – cyber bullying. It is a form of harassment that occurs via Internet which involves vicious forum posts, name calling in chat rooms, creating fake profiles, and mean email messages. As the number of social media users tremendously increases, it consequently intensifies the cyber bullying problem. Legal and ethical issues are now confronting the abuse of using the social media to bully other internet users. This is why the development of automated systems is essential to help monitor activities occurring in the cyberspace. Successful detection of cyberbullying occurrences in the Internet is therefore of key importance to identify and classify possible threatening situations online and prevent them from escalating.

## Business Opportunity

The target market of the present study are system developers and researchers. Since the team will be constructing a cyberbullying detection model, it will capture the interest of the researchers who want to explore the features of Natural Language Processing and Support Vector Machine (SVM) algorithm. As for the system developers who want to create an automated cyberbullying detection system, this study will provide them a model that they can integrate into social networking sites to monitor the activities in the cyberspace automatically.

As cyberbullying becomes rampant: authorities, system developers and researchers proposed and implemented different solutions to mitigate this problem. These methods involved manual review tasks of identifying offensive contents, integrating lexicon-based approach to filter offensive contents, communication model to track online predators, lexical features with machine learning classifiers to determine the victim and a bully in each instance of cyberbullying, profiling features to detect aggressive discussions and others. However, some of these methods were proven to be inefficient because they rely merely on the user to report cyberbullying incident before they take actions. It is also not possible for the moderator to detect all cyberbullying occurrences in social networking sites. The automatic offensive content detection, on the other hand, often yields low accuracy, slows down the system processing speed, and are prone to more false positives. To address this limitations, the team decided to create a cyberbullying detection model that can generate up to 70-80% accuracy in detecting vicious online contents.

## Business Objectives and Success Criteria

The research aims to formulate a cyberbullying detection model which will yield at least 70-80% accuracy in terms of detecting cyberbullying occurrences present in public social media posts through text classification using Support Vector Machine (SVM) approach. As soon as the Support Vector Machine algorithm has been implemented to the cyberbullying detection model, it can now be used by the target market.

## Customer or Market Needs

A current research that was conducted concludes that 8 out of 10 Filipinos are using social networking sites. With the rapid growth of social media, Filipino users are spending significant amount of time on different social networking sites to establish connection and relationship with peers. While they benefit from their use of social media by interacting with others, social media users are also at the risk of being exposed to a large amount of vicious online contents. Therefore, cyberbullying in the Philippines has become severe. This led to the introduction of Anti Cybercrime Act of 2012, which recognizes cyberbullying as a kind of cybercrime and provides provision on the consequences for cyberbullying. However, despite the law, it is not possible to monitor the Internet constantly for any acts of cybercrime being committed. Another method was used in Japan by the Parent Teacher Association (PTA) wherein they perform website monitoring called “net-patrol”. Once a harmful post was detected by a net-patrol member, he will immediately report it to the administrator. However, manual monitoring of offensive content in social media has proven to be inefficient, time consuming, and they are not scalable in reality. This is why automated systems must be developed to help monitor activities in social networking sites and to lessen the efforts of moderators.

Several methods have been implemented as well to detect cyberbullying occurrences automatically. In order to detect harmful entries, techniques such as offensive content filtering methods in social media and text mining techniques have been implemented. Popular social networking sites apply several mechanisms to screen offensive contents. To illustrate, Youtube’s safety mode, once activated, can hide all comments that contains harmful content from users. And on Facebook, users can add comma separated keywords to the “Moderation Blacklist” so when people include blacklisted keywords in their post or comment, it will be automatically identified as spam. Most social networking sites use lexicon-based approach to filter vicious posts. Furthermore, they rely solely on users to report offensive contents to take actions. These systems are proven to have low accuracy and may generate many false positive alerts due to their use of simple lexicon-based automatic filtering approach to block harmful posts and comments. While defensive methods adopted by social networking sites were inefficient, researchers have proposed different ways to identify offensive context using text mining approach: moreover, implementing text mining techniques in analyzing online data requires the following phases: data acquisition and preprocess, feature extraction and classification. Thus, the major challenges of using this technique lie on the feature selection phrase. Most automatic detection systems extracts two kinds of features: lexical and syntactical features. Lexical features treat each word and phrase as an entity. Word patterns such as appearance of certain frequencies are often used to represent the language model. However, it fails to distinguish sentences’ offensiveness which contain same words in a different order. On the other hand, syntactic features use natural language parsers to parse sentences on grammatical structures before feature selection. While these methods only focused on sentence-level and message-level constructs, there were also some limited efforts at the user level. However, it is a more challenging task. To illustrate, Kontostathis et al propose a rule-based communication model to detect and classify cyberbullies. Pendar uses lexical features with machine learning classifiers to differentiate victims from predators in online chatting environments. Pazienza and Tudorache propose using utilizing user profiling features to detect aggressive discussions. However, despite the continuous efforts in incorporating user information in detecting harmful online contents, more advanced user information such as user’s writing styles and posting trends has not been included to improve the accuracy of an automatic detection system.

In order to create a cyberbullying detection model, the researchers used the following tools: Import.io, a web scraper that was used in extracting data from the web. Weka, an open source software that performs pre-processing techniques. Excel and Notepad++ for data normalization and removal of unnecessary special characters. Lastly, Brat and GATE, these are open source software that were used for data annotation.

The cyberbullying detection model can be integrated by the software developer to social media in automating the process of cyberbullying detection. As for the researchers, through the use of the manual or project documentation, they can gain insights about Natural Language Processing and Machine Learning Algorithm which they can use as their guide for their research.

**1.5 Business Risks**

## Although the cyberbullying detection model is based on Philippine setup, it can only detect Tagalog and English words which contain indications of cyberbullying. Native languages from the provinces are outside the scope of this project. Furthermore, cyberbullying roles will not be identified as well only the classification of cyberbullying statements into sensitive issues.

# Vision of the Solution

## Vision Statement

The team’s proposed output for the research that they are currently working on is a cyberbullying detection model. This model will be able to give computer systems the ability to analyze public social media posts according to predefined patterns and subsequently, detect cyberbullying occurrences present within them (if there are any). It is expected to provide an accuracy rate of at least 70-80% in terms of flagging such words and classifying them into classes (which will likewise imply the reason as to why they were flagged in the first place). Its main purpose is to give researchers an idea on how to extend the process of social media moderation as it is currently being done here in the Philippines - from completely involving human intervention into becoming more automated – at least in terms of the detection of the statements of interest. Additionally, interested developers can also use this particular model alongside their software creations in order for them to build something that can actually make a difference among Filipino netizens. In the event that the incorporation of the model with a software becomes possible, social media moderators can now be warned of imminent cyberbullying occurrences (as soon as the words were detected) which can therefore allow them to administer appropriate actions before the situation gets out of hand. This is because the detection process will not solely be based on the results of a word-for-word matching technique. Instead, the model will also consider the meaning and context of a particular word or group of words when combined together (as it was proven by another study that there are words which would only bear negative connotations when used together with other words in a sentence). Most importantly, it will greatly help Filipino netizens, especially those who are afraid of reporting such statements to the designated authorities and those who are not even aware that the statements which were addressed to them contain implications of a probable cyberbullying attack, in providing assurance for their own safety and security in the online world, and giving them a chance to stand for themselves against the bullies.

## Major Features

1. Modeled according to the context of cyberbullying in the Philippine setup (specifically in Metro Manila)
2. A corpus that contains statements in Filipino (which uses Tagalog – as how it is conventionally spoken in Metro Manila - and English words).
3. All statements were annotated mostly by adolescents – the age group which utilizes social media more frequently compared to the rest according to a recent study.
4. Has an accuracy of at least 70-80% in terms of detecting cyberbullying instances and classifying them in accordance to the reason as to why they were flagged

## Assumptions and Dependencies

The main assumption that the researchers did in terms of measuring the success of the project is by giving an accuracy rate of at least 70-80% despite the fact that they themselves are not aware of how much will they have to tweak in their model in order to reach the said threshold. Additionally, they likewise assumed that automated social moderation systems are not conventional here in the Philippines based on the amount of research papers that they gathered with regard to the said topic. The algorithm tests will be heavily dependent on the algorithms that the WEKA toolkit contained. Although they planned to pursue other algorithms (those unconventional ones), they are still trying to limit it to the point where doing such would not give them a difficult time.

# Scope and Limitations

## Scope of Initial Release

The cyberbullying detection model will be developed and experimented against factors which are considered as the main constituents of cyberbullying scenarios in the Philippines, specifically in Metro Manila only. Statements that were used to populate the corpus were extracted from social media website contents which many Filipinos deemed controversial (e.g. regarding inappropriate behavior, gender, and the like). The researchers believed that if they were to get their statements on the said areas, they will be able to get the gist of what most Filipinos refer to as “offensive”. About 2000 statements extracted from these posts will be used in the entirety of the corpus. On the text pre-processing phase, the annotation of the statements (considered as cyberbullying statements) will be based according to the different categories taken from the controversial issues. The said annotations will likewise serve as the classes that will be utilized during the feature extraction stage. The team also considered the following text preprocessing methods (apart from the text annotation process) as of this moment: tokenization and removal of extra characters. The feature extraction stage, which follows after the text pre-processing phase, will involve the use of an algorithm known as Bag-of-Words (BoW) to extract the features in every category (or class). These features will then become the basis for the model’s performance during the testing stage. The adjustments that will be made (e.g. the possible removal or inclusion of more features in the training data) against Support Vector Machine (SVM) algorithm will take place in the experimentation stage.

In the initial release, the cyberbullying detection model will remain intact with WEKA toolkit and will be presented in that way for the sole purpose of demonstrating how the detection process would work.

## Scope of Subsequent Releases

In the subsequent releases, the team is planning to integrate the model with an external application (a Java API) in order to provide a basic Graphical User Interface for it. It will once again showcase a group of statements (social media posts extracted from various social networking sites) which will be assembled in such a way that it can provide a clear view of what the project’s stakeholders’ should expect on how the model would work in the background (in the event of its integration with a particular software development).

## Limitations and Exclusions

Since the group had agreed early on that the only output that they will present along with their research paper is the cyberbullying detection model alone, one must not expect a software (of the same level as that of a plugin designed to cater to social media moderation – complete with user-friendly GUI and optimized views) development of some sort to be included as proof of the model’s accuracy in doing its expected task.

# Business Context

## Stakeholder Profiles

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Stakeholder** | **Major Value** | **Attitudes** | **Major Interests** | **Constraints** |
| Researchers (the team itself/the students creating this research paper) | Proves the feasibility of the project | Views the creation and successful implementation of the cyberbullying detection model as their main objective | To be able to put their research into good use (in terms of creating an impact on the lives of other people, especially their fellow countrymen) | Expertise in the field of Natural Language Processing and the time that they can only a lot in the project-making process |
| Other researchers (interested in the field of technology) | Extends the capabilities of the research paper and may share ideas regarding the contents that are to be included | Views the paper as a source of reliable data to add to their current knowledge in terms of automated cyberbullying detection systems modeled following the Philippine setup | To be able to develop the paper into a more profound document which would encourage the idea of technological innovations to become widespread in the Philippines | Availability and extent of the resources he/she is only willing to share |
| Computer Science students (undergraduates) | To serve as basis for their own research paper (if inclined towards the same topic) | Do the best that they can in trying to analyze the concepts needed for their research paper | Acquire the ability to understand such concepts and search for ways to learn more stuff related to their course | Lack of knowledge in the said field (may lead to the paper to turn out as something that is very basic) |
| Graduates of Computer Science | Can be of help in terms of providing explanation with regard to processes or terms in the field of Computer Science | Tend to show interest in the concepts and processes involved in the said field | Wants to see for themselves how Computer Science undergraduates are currently being equipped with the important concepts in the course | Availability and may not be good in sharing ideas to beginners |
| Experts in Psychology | Contributed in defining cyberbullying in the Philippines | Tend to be enthusiastic and approachable in dealing with students who sought advice from them | Determine peoples’ behavior under certain circumstances | Lack of expertise in the field of technology |
| Developers | Can turn the purely conceptual cyberbullying detection model into a component of a software development | Focuses mainly on the technical aspects of the system | Search for new ideas which would give them an edge in the market | Is not particular with the conceptual aspect of the research |
| Project Adviser | Looks for connections (or people who can contribute to the entirety of the research paper) | Maintains good connections with different people in both the industry and the academe | To make the team’s research project profound and renowned by offering to help in the authoring of the entire paper | Availability and time spent on the project |
| Project Consultant | Guides the researchers as they undertake their project-making process | Encourages the researchers to finish creating their project and provides ample advice as to how they should deal with certain scenarios that they may face in the process of creating their project | Taking note of the students’ growth and development during the process of undertaking their project and encourages them to submit/publish their research paper in both local and foreign research symposia | Is typically dedicated to another work as well |

## Project Priorities

|  |  |  |  |
| --- | --- | --- | --- |
| **Dimension** | **Driver (state objective)** | **Constraint (state limits)** | **Degree of Freedom (state allowable range)** |
| Schedule | Release 1.0 to be available by the end of the third term of A.Y. 2016-2017, release 2.0 by the end of the first term of A.Y. 2017-2018 | Commitments/requirements from other subjects that must also be met/accomplished | At least the first half of version 1.0 must be finalized at the end of this term (in order to meet the stated ideal schedule of project version releases) |
| Features | Detect cyberbullying occurrences in public social media posts written in Filipino | Must include English and Tagalog words that are commonly used or can easily be comprehended by most people living in Metro Manila | The detection and classification of detected cyberbullying occurrences in the social media posts must yield an accuracy rate of at least 70-80% as much as possible |
| Quality | Present a clear interpretation of how the model is expected to work (with a system) with the help of a prototype | The prototype will adapt a simple GUI (one that is enough to show how the detection process would work using the developed model) | 80-85% of user acceptance tests must pass for release 1.0, 90-95% for release 2.0 |
| Staff | Must exercise teamwork and constant communication among each member | Maximum of 3 people (in every group) + 1 adviser + maximum of 5 consultant | Can allow up to 4 in the event of accepting other individuals who have yet to belong in a particular group |
| Cost | Must trim down costs to values as low as possible | Only incur costs when emergency arises (e.g. additional resources that the group has to acquire for a fee) | At least 90% of the costs incurred must be deducted from each members’ capstone fee (or at least until it still has not run out) |