Lameck Onyango Oyare

MSc. Software Engineering

Reg.SCT313 – 3506/2017

**Continuous Assessment**

Implementing an empathic agent to detect textual Cyberbullying

# 1.0 Introduction

The social nature of Web 2.0 applications is increasingly affecting communication and collaboration in today’s society. It was argued that message boards, blogs and social networking platforms like Facebook, Twitter, telegram, Instagram or WhatsApp have become an important means of communication and collaboration, especially among teenagers (Van Hee, C). According to Soko Directory (2017), a total of 12 million Kenyans were on WhatsApp with 7.1 million of them on Facebook. Although most of the time, youth’s Internet use is thought to be perfectly safe and enjoyable, there are risks involved with the Internet on social media websites. Like offline communities, online communities can be harmful. Youngsters can be confronted with threatening situations, such as cyberbullying, suicidal behavior or grooming by pedophiles.

As a response to those threats, governments have come up with Cybercrime laws as preventive initiatives. For instance, the Kenyan government enacted the Computer Misuse and Cybercrime Act to increase online youth safety. In spite of these efforts, much undesirable or even hurtful content remains online.

According to nderi, s. (2017, September 27) Cyberbullying can be described as the use of cell phones, instant messaging, email or social networking sites such as Facebook and Twitter to harass, threaten and intimidate someone. Back in the day, bullying used to be on the playgrounds of schools or in a case of high school in the dorms where the older boys used to pick on the formative. However, with the advent of technology, bullying has gone digital whereby someone can bully you anonymously and from anywhere in the world. All one needs is access to communication technology. Given the gravity of the problem and its rapid spread among the youth, there is an immediate and pressing need for research to understand how textual cyberbullying occurs today so that techniques can be rapidly developed to accurately detect, prevent, and mitigate textual cyberbullying.

Sentiment analysis is generally denoted as techniques used to determine the predisposition of text, usually expressed in free text form. In this research, I used sentiment analysis for text classification and analyze incoming messages and tell whether the underlying sentiment is positive or negative.

Subjective information in source materials is recognized and extracted by the means of natural language processing, text analysis, and computational linguistics. It is used to determine an author's attitude, with respect to a particular topic or the overall contextual polarity in the text. This is a promising technology, which has resulted in remarkable interest among academics. The development of research in the field of text analytics has allowed researchers to formulate algorithms and techniques to discover sentiments from free text more efficiently than ever. This work covers all features from mining texts from social media, applying sentiment analysis based on people’s opinions expressed on social media to finally assigning the polarity to them as positive, negative or neutral.

## 1.2 Problem Statement

The word cyberbullying did not even exist a decade ago, yet the problem has become a pervasive one today. Cyber bullies do not have to be strong or fast; they just need access to a cell phone or computer and a desire to terrorize. Anyone can be a cyberbully, and such persons usually have few worries about having face-to-face confrontation with their victims.

In this paper, we focus on the detection of textual cyberbullying with an empathic agent, which is one of the main forms of cyberbullying. We use a corpus of comments from YouTube videos involving sensitive topics related to race & culture, sexuality and intelligence i.e., topics involving aspects that people cannot change about themselves and hence become both personal and sensitive. We pre-process the data, subjecting it to standard operations of removal of stop words and stemming, before annotating it to assigning respective labels to each comment.

According to the survey by Digital Trends, more youths experienced [cyberbullying on Instagram](https://www.digitaltrends.com/mobile/instagram-anti-abuse-tools/) than any other platform at 42 percent, with Facebook following close behind at 37 percent. Snapchat ranked third at 31 percent. Seventy-one percent of the survey participants said that social media platforms do not do enough to prevent cyberbullying.

The survey also considered the other side of the story, asking the same age group how often they were the bullies, instead of being on the receiving end. Nearly 70 percent of those surveyed said they were abusive online toward another user, compared to just 12 percent that admitted to bullying in general. Despite the prevalence of youth initiating the bullying, more than 60 percent disagreed with the idea that “saying something nasty” is less hurtful online than in person.

## 2.2 Related Works

Many researchers have worked mostly with finding out crime pattern from social media or Internet blog using sentiment analysis. Very few researches have been conducted on the context of cyberbullying. Their approach differs from ours in many ways and so; we will discuss a brief summary of their research and results.

Kenya recently enacted a bill the Computer and Cybercrimes Bill (2016) whereby harassing and stalking someone on Facebook or Twitter can now earn you a 10-year prison sentence or a Sh20 million fine or both. This follows Cabinet’s approval of the Computer and Cybercrimes Bill (2016) that spells out stiff penalties for digital crimes including illegal breach of systems and networks, cyber-bullying and stalking among others. “A person who, individually or with other persons, willfully and repeatedly communicates, either directly or indirectly, with another person or anyone known to that person, commits an offence, if they know or ought to know that their conduct is likely to cause those persons apprehension or fear of violence to them or damage or loss on that persons’ property detrimentally affects that person,” reads Section 14 of the Bill in part.

The Computer and Cybercrime Bill 2016 is part of a raft of legislation mooted by the Government by various agencies in the last three years to combat rising cases of cybercrime.

Previous psychological and sociological studies on the bullying behaviors and emotional intelligence suggest that emotional information can be used to better understand the bullying behaviors. Emotional intelligence refers to the ability of an individual to accurately perceive emotion, use emotions to facilitate thought, understand and manage the emotion. The lower the emotional intelligence of the user, the more likely an individual will be involved in the bullying behaviors. Motivated from this insight, we investigate if the use of sentiment information of the post content could help better understand and accurately detect cyberbullying behaviors in social media.

# 2.0. System prototype

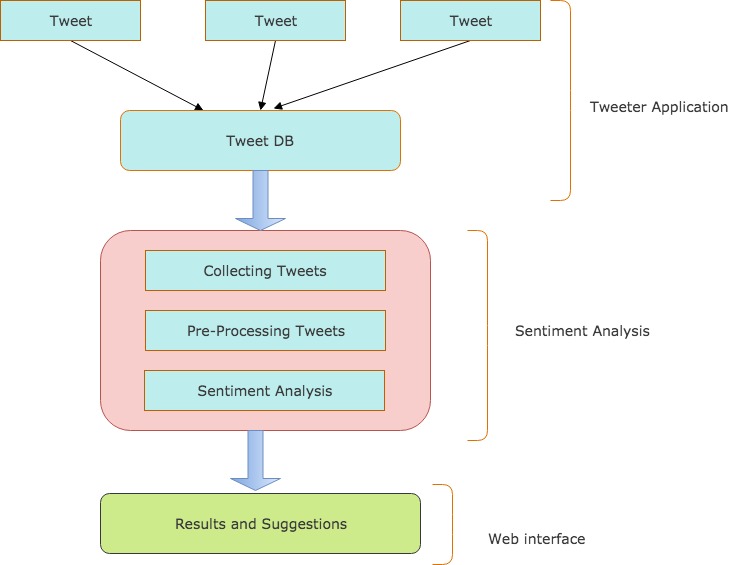
## 2.1 System Design

**Figure 2.1** shows the basic architectural diagram of the implemented system. Basically it consists three modules, they are:

• Twitter application and twitter database

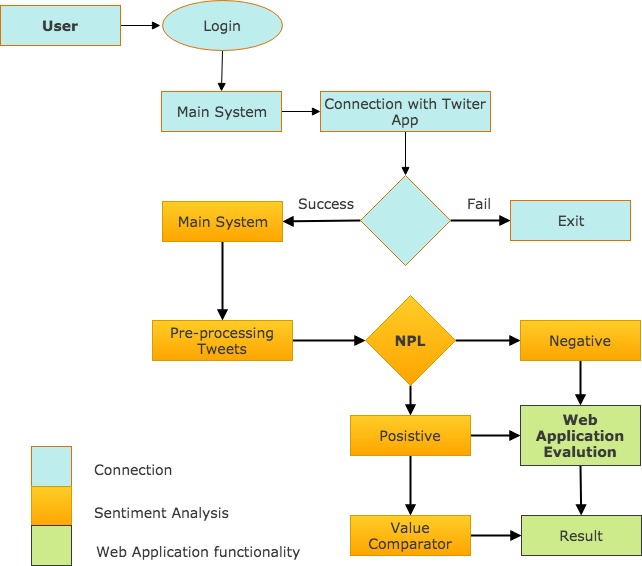
• Sentiment analysis process

• Web application



**Figure 2.1 architectural designs**

**Figure 2.2** shows the dataflow diagram of the implemented system. At first, a twitter application is created and tweets are collected from the twitter database. Collected tweets are stored as data set and is pre-processed and parsed by removing common unwanted words, symbols, characters, numbers and converts the upper case letters to lower case letters. After pre-processing, the sentiments will be analyzed by using Natural language processing tool. Each sentence is provided with sentiment value, based on this sentiment value the data is cataloged as positive or negative. We will also be able to detect cyber bullying on the text that has been classified as negative by Natural processing Language. Both positive and negative data are analyzed and similar data are identified. Then by using a web application,



**Figure 2.2 Data Flow Diagram**

## 2.2 Data Collection:

This is happens on the connection part. I used twitter to collect relevant data since it is one of major social networking website. To get access and gather I developed an API with python that integrates with twitter to pool all the tweets. You first needed to login then the API establishes a connection with the twitter app to gather all the data required then move it to the database within the main system.

## 2.3 Pre- Processing:

Initially for pre-processing I corrected the spelling mistakes in the tweets as many people tend to right English words in short form i.e., “u r nt pretty” is converted to “you are not pretty”. Then we converted the uppercase letters of the tweets to lowercase order. Then we removed all the usernames, URLs and unnecessary white spaces from the tweets. Stop words are words that are generally considered useless. Most search engines ignore these words because they are so common that including them would greatly increase the size of the index without improving precision or recall. Any group of words can be chosen as the stop words for a given purpose.

For some search engines, these are some of the most common, short function words, such as the, is, at, which, and on. In this case, stop words can cause problems when searching for phrases that include them, particularly in names such as " The Who ", " The The ", or " Take That ".

## 2.4 Natural Language Processing

Using natural language sentiment analysis requires a set of tweets labeled positive, negative or objective, it can be created by hand, by labeling tweets manually. I opted for a more automatic approach by collecting hundreds of thousands of tweets and running an algorithm through those tweets, which compares each individual word with positive list and negative list of words.

## 2.5 Evaluation

The tweets that are positive of cyber bullying are flagged and passed onto value comparator for further screening. While the negative ones will be displayed on the web interface as a normal tweet.

**References**

1. Fabian Fagerholm, Scientific writing , (2016) online . Available on https://www.cs.helsinki.fi/group/ese/ScientificWritingGuide.pdf
2. soko directory, (February 15, 2018) , online https://sokodirectory.com/2018/02/number-kenyans-social-media-platforms-increases-bake/
3. nderi, s. (2017, September 27). Cyber bullying as a form of violence against women & how to tackle it. Retrieved from <https://hapakenya.com/2017/09/27/cyber-bullying-as-a-form-of-violence-against-woman-and-how-to-tackle-it/>
4. Kenya cyber crime act
5. GitBook, Sentiment Analysis, online: <https://lizrush.gitbooks.io/algorithms-for-webdevs-ebook/content/chapters/sentiment-analysis.html>
6. UML diagrams, online: <https://www.draw.io/>