**Implementing an empathic agent to detect textual Cyber bullying**

**Introduction**

The social nature of Web 2.0 applications is increasingly affecting communication and collaboration in today’s society. It was argued that (Van Hee, C) Forums or message boards, blogs and social networking platforms like Facebook, Twitter, Tumblr or WhatsApp have become an important means of communication, 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 perfectly safe and enjoyable, there are risks involved in online communication through social media. Like offline communities, online communities can be harmful. Youngsters can be confronted with threatening situations, such as cyberbullying, suicidal behavior or grooming by paedophiles.

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. Inspite of these efforts, much undesirable or even hurtful content remains online.

According to whatIs.com Cyber bullying 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 the case of high school in the dorms where the older boys used to pick on the form ones. 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 will use sentiment analysis for text classification and analyze incoming messages and tell whether the underlying sentiment is positive or negative.

**Literature Review**

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.

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.

In this paper, we attempt to perform cyberbullying detection in a supervised way by proposing a principled learning framework. More specifically, we first investigate whether sentiment information is particularly correlated with cyberbullying behaviors. Then, we discuss how to deal with short, noisy, unstructured content and how to properly leverage sentiment information for cyberbullying detection. Methodologically, we present a novel learning framework called Sentiment Informed Cyberbullying Detection (SICD). Experiments on two real-world social media datasets validate the effectiveness of the proposed framework.