Jomo Kenyatta University of Agriculture and Technology

Dept: Telecommunication and Information Engineering.

CYBERBULLYING DETECTION USING SENTIMENT ANALYSIS

PROJECT PROPOSAL

Presenter

BERNICE WAWERU ENE221-0277/2016

CYBERBULLYING DETECTION USING SENTIMENT ANALYSIS

BY

BERNICE WANGUI WAWERU

ENE 221-0277/2016

A project report submitted to the Department of Telecommunication and Information Engineering in fulfilment of the requirements for the award of the degree of Bachelor of Science in Telecommunication and Information Engineering at the Jomo Kenyatta University of Agriculture and Technology.

19th November 2022

Declaration of authorship

This project report is my original work and has not been presented for a degree in the University.

Signature ……. Date ……19th November 2022

Name ……BERNICE WAWERU. Registration number ENE 221-0277/2016

Supervisor's approval.

This project report has been submitted for examination with my approval/knowledge as university supervisor.

Signature ……. Date …….

Name …….

Acknowledgments

My sincere thanks to my supervisor for his invaluable input and to my friends and family for supporting my efforts. I give thanks to God for his provision and care.

Table of Contents

List Of Figures

ABSTRACT

The internet has increased connectivity and opened new opportunities for people to communicate through different methods, including social media, blogs, websites, and emails. Unfortunately, some people misuse these platforms to harass and intimidate others, creating a hostile cyberspace. Cyberbullying has been on the rise in recent years as the internet has become more accessible due to greater penetration in different places and the availability of devices such as mobile phones. More people are susceptible to online bullying due to the large audience that can contribute to trolling an individual based on one bully's comment. The perpetrators can also remain anonymous; thus, bullying can get out of hand because the individual knows they are not likely to face any consequences. Victims are more likely to develop anxiety due to the trauma of bullying and may also develop depression in extreme situations. Cyberbullying is a public health issue that should be addressed to mitigate its effects and ensure people feel safe using different digital communication platforms.

This project aims to build a service that can be used on different online platforms to analyze sentiments and determine whether the sentiments expressed are intended for bullying. The Bidirectional Encoder Representations from Transformers model, considered a state-of-the-art natural language processing model, will be used for sentiment analysis. The project will focus on detecting bullying based on the text by deriving the feelings expressed. The project is designed to be used by other services that analyze users' sentiments.

Introduction

Background

The availability and accessibility of the internet in various countries have facilitated the rise of social media platforms and the use of different forms of electronic communication. For instance, the Internet World Stats (IWS) indicates that Kenya has an internet penetration rate of 84.1% as of December 2021, demonstrating its prevalence and ease of access for most citizens [1]. In addition, Kenyans have adopted social media platforms significantly, with Facebook, Twitter, and WhatsApp being the most prevalent [2]. The internet and social media platforms offer several advantages that have contributed to their popularity and widespread adoption by individuals and organizations. Some advantages include reaching a broader audience, improving communication and interactions among individuals, and opening new trade and cultural exchange opportunities.

Unfortunately, despite the numerous advantages of using social media, it has become a breeding ground for harassment and bullying as perpetrators use the platforms to intimidate and abuse victims. Cyberbullying has become rampant as more people use social media to voice their opinions and seek support from people who subscribe to similar schools of thought. The freedom and ease of sharing opinions make social media rife for cyberbullying, as people can share threatening and abusive opinions about others and spark similar reactions from their networks. Additionally, the lack of clear guidelines on the consequences of cyberbullying allows people to bully others without worrying about being punished because most countries have not adopted any laws to deal with such scenarios.

The coronavirus pandemic led to more organizations and individuals adopting online platforms as the preferred way of conducting business, increasing the number of people susceptible to bullying. Children and young adults are vulnerable to harassment online as they interact on various platforms for educational or entertainment purposes. Companies also use dedicated internal communication software, which presents opportunities for bullies to spread workplace harassment and cyber victimization. Celebrities and other famous personalities are often trolled and bullied online for their opinions or based on the content they publish online. It is vital to acknowledge that even ordinary people can also be victims of cyberbullying, as demonstrated by a study by the World Wide Web Foundation, which contends that one in every five women in Kenya has experienced cyberbullying [3].

Cyberbullying severely impacts the victim's well-being and can have long-lasting effects on the person's interactions; thus, it is vital to prevent bullying and mitigate its effects. Advances in Natural Language Processing (NLP) and machine learning offer a solution to detect negative sentiments that may be abusive or threatening to a person. Pre-trained models in computer vision have demonstrated remarkable performance triggering similar practices in NLP to improve the accuracy of various tasks. This project will use the Bidirectional Encoder Representations from Transformers model, a pre-trained model trained on about a 250million words and with demonstrable performance improvement. The model was initially used at Google to improve search accuracy but was later made open source making it available for use for different use cases. BERT is highly context-aware due to the bidirectional technique used where it derives context from the left and right of a given phrase; thus, it can determine the original context in the sentence. Context is essential in predicting cyberbullying because some phrases can be offensive or harmless depending on context.

Project Description

The project will rely on sentiment analysis and natural language processing techniques to derive context and classify the texts and posts accordingly. The proposed system will leverage the advancement of machine learning and the availability of enormous amounts of data to train a model that can classify and detect different forms of cyberbullying by analyzing texts. The system will be an application programming Interface (API) that other clients, such as social media platforms and companies' internal communication tools, can consume. The clients send a request containing the information they need to be analyzed, and the services return a response containing details on whether the sentiments expressed in the request data are intended for bullying.

Purpose

This project aims to decrease cyberbullying by facilitating the early detection of hurtful, abusive, or threatening language toward an individual**.** This project is designed to enable other services to rely on it by making calls to the API, which sends a response payload that includes the classification of the text.

Objectives

The objective of this project is to design and implement a system that will:

- Detect bullying by identifying the threatening language in comments
- Detect bullying by identifying the abusive language in comments
- Return classification of messages to the system's clients.

Literature Review

Cyberbullying

The United Nations Children's Fund (UNICEF) identifies cyberbullying as any form of harassment perpetrated through digital technologies such as social media, messaging platforms, and gaming platforms [4]. Bullying online can take different forms, including phone calls, text messages, photos or videos and emails, blogs, chat rooms, and other forms of instant messaging [5]. Ferrara et al. point out that cyberbullying is characterized by the lack of direct contact between the victim and the perpetrator and the possible anonymity of the bully [5]. The anonymity often increases the rate of aggression as abusers can hide behind their keyboards without worrying about direct consequences. Bullying online also has the potential to reach wider audiences, increasing the number of bullies that a victim has to face. Moreover, it is often hard to disconnect themselves from the cyber environment despite being bullied because most communication nowadays is done online.

Traditional bullying is associated with violence, humiliation, and intimidation of the victims during their physical interactions with the perpetrators in school or the office. Although cyberbullying has some similarities with traditional bullying, it has differences that exacerbate its impact on the victim. Bullying online can happen at any time, and individuals are susceptible to harassment, unlike traditional bullying, when the victimization seizes once the victim goes away from the abuser. Zhu et al. [6] posit that cyberbullying is a significant public health issue affecting children and young adults. Although cyberbullying may often be mistaken as making fun of a person without malicious intent, it can affect the person severely on a larger scale.

Bullies online have greater access to their victims and can cause significant damage using the information they find online and create false information. Ferrara et al. state that cyberbullies can use personal information for identity theft, hacking the victim's computers or devices [5]. The harassers can also impersonate their victims and create a false impression of the person damaging their reputation by posting embarrassing content. In addition, bullies can also orchestrate scenarios that create an environment for further harassment by sharing the victim's personal information, such as their address or phone number. In other environments, bullying may be through the exclusion of the victim in online activities such as meetings and groups where their contribution is expected. Such exclusion is meant to humiliate the person because they miss important information and context necessary for other interactions. Shearman asserts that exclusion is one of the most prevalent forms of online bullying in the workplace, which is a passive-aggressive move that can be disregarded and dismissed as a mistake [7].

Abaido asserts that the victims rarely report incidences of cyberbullying because they feel helpless [8]. Children and adolescents may also be hesitant to report bullying because they fear their parents or guardians will take away their devices. The psychological harm caused by cyberbullying supersedes that inflicted by traditional bullying because there is a larger audience, and the content can be preserved and used to traumatize the victim repeatedly [8]. Effects of cyberbullying include suicidal behaviour, anxiety, depression, substance abuse, sleeping and eating disorders, and decreased academic performance among students [9]. Employees' morale, performance, and productivity are also negatively affected by online aggression. It is vital to protect people from bullying by creating a safe online environment where individuals express their opinions without harassing or threatening others.

Sentiment Analysis

Sentiment analysis refers to gathering and analyzing people's opinions, thoughts, and impressions regarding various topics, products, subjects, and services [10]. Sentiment analysis is also known as opinion mining and involves determining the feelings expressed by people about a topic through different posts and other reactions that various platforms allow them to use to express themselves [11]. Sentiment analysis also classifies the post's polarity into different categories, such as positive, negative, or neutral. With the rise of social media platforms and interactive websites that allow users to leave feedback, sentiment analysis has become one of the most common ways to understand customers' opinions. Wankhade et al. point out that sentiment analysis has been adopted in businesses, governments, and other organizations mainly for customer reviews [12]. Some everyday use cases include analyzing product reviews for e-commerce platforms, analyzing hotel customer reviews, and software application reviews.

Alaoui et al. identify that sentiment analysis can be divided into lexicon analysis and machine learning [10]. Lexicon analysis calculates the polarity of a text from the semantic orientation of words or phrases in the document. Machine learning (ML) involves building models that can predict the polarity of a text or post based on training models using the training dataset. Sentiment analysis utilizes Natural Language Processing (NLP) to extract, convert and interpret opinions from a text and classify sentiments [12]. Natural Language Processing is an emerging field in machine learning that focuses on facilitating the ability of computers and systems to understand and analyze human language in spoken or written form. The field has seen significant advances that have led to the adoption of various NLP techniques to create robust systems, such as speech recognition software.

Sentiment analysis can be done on several levels depending on the intended use case. It can be done on the document, sentence, phrase, and aspect levels [12]. The entire document is analyzed at the document level to determine its general polarity. In contrast, at the sentence level, individual sentences are considered to identify their polarity. Phrase-level sentiment analysis forms the foundation for most sentiment analysis tasks because it is conducted on phrases, and their sentiments are classified. Aspect-level analysis can be done for sentences with multiple aspects. The availability of large datasets from social media platforms and other websites has contributed to NLP and ML improvements. The emergence of techniques such as deep learning has helped improve the performance of ML models, increasing confidence in their performance; thus, more companies are willing to deploy their machine learning projects to production. The models' performance is improved based on real-world and real-time data; therefore, training, validating, and testing models is an iterative process.

Several researchers have used sentiment analysis for cyberbullying detection, implementing different machine-learning techniques to build models that predict whether the text counts as harassment. Almutiry and Fattah used the Support Vector Machine (SVM) to conduct sentiment analysis in Arabic and determine whether posts were meant for bullying [13]. Recurrent and convolutional neural networks are also some of the deep learning techniques used for NLP tasks and cyberbullying detection [14]. In addition, available literature on cyberbullying detection focuses on determining the performance of various models on the same dataset. For instance, Rosa et al. compared the performance of different ML algorithms in detecting bullying by comparing Random Forests, Support Vector Machines, and Logistic Regression [14].

Currently, one of the most widely used and practical models for NLP tasks is the Bidirectional Encoder Representations from Transformers (BERT) model, introduced in 2018 by researchers from Google AI Language [15]. The model has demonstrated its capabilities in NLP tasks where one can reproduce high accuracy levels even with a small training dataset. BERT derives its advantages from bidirectional training of the transformer, a popular attention model used in language modeling. Unlike other methods used in training that rely on analyzing text from left to right or combined left-to-right and right-to-left training, BERT considers the text to the right and left of a given phrase. Bidirectional-trained language models develop a better understanding of the language context and its flow and thus perform better in NLP tasks [15]. This project uses the BERT Transformer to predict harmful sentiments leveraging the model's additional performance and capabilities.

Bidirectional training uses Masked Language Model (MLM), a novel technique that has facilitated bidirectional training that was previously impossible [14]. In MLM, some random masks are used on some input tokens, and the model predicts the original masked phrase based on context. BERT transformer was developed based on the premise of transfer learning and pre-trained models, which have gained traction in recent years, especially in computer vision; thus, researchers have adopted similar techniques in NLP. Transfer learning refers to pre-training a neural network model on a known task and then fine-tuning it to meet the expected need [15]. Transfer learning improves the performance of a model because it relies on knowledge learned from previous tasks to solve the current problem. Devlin et al. assert that pre-trained BERT models can be used for various tasks by fine-tuning the output layer because it is designed to pre-train deep bidirectional representations from an unlabeled text by conditioning on the left and right context in all layers [15].

BERT builds upon transformers, a common self-attention mechanism in deep learning that derives context from weighing the significance of each part in the input data [16]. A transformer uses an encoder-decoder mechanism where the encoder reads the text input, and the decoder predicts the output. Gillioz et al. identify that transformers are attention-based NLP techniques that counter the limitations of Recurrent Neural Networks, which are sequenced to sequence-based models with a significant challenge when dealing with long-range dependencies [17]. Models can implement different types of attention depending on where it is placed. Attention between the input and output sequence is called encoder-decoder attention, while self-attention can be used in the input or output sequence. BERT transformer uses self-attention in the input sequence, and the transformer encoder reads the entire sequence of words simultaneously. A sequence of tokens is used as the input embedded into vectors and then processed in the neural network to get a sequence of vectors as the output.

BERT uses Masked Language Modelling and Next Sentence Prediction strategies to overcome the challenges in context learning. Masking is done by replacing 15% of the words in each sequence with a [MASK] token, and then the model attempts to predict the original word in the masked token [15]. It relies on the context of the words preceding and after the masked token. The Next Sentence Prediction technique involves giving pairs of sentences to the model and training it to predict whether the second sentence is an appropriate subsequent sentence. These approaches ensure that the BERT transformer is context aware thus, it is very suitable for cyberbullying detection where context is vital because people use language differently, and the implied meaning changes over time.

Several studies have used the BERT model for various natural language processing tasks because the model can be fine-tuned to achieve the desired sentiment classification. Bilal and Almazroi used BERT to classify online customer reviews determining whether the reviews were helpful or not [19]. Their research indicates that the fine-tuned BERT-based classifiers outperform bag-of-words approaches, including the K-Nearest Neighbors, Naïve Bayes, and Support Vector Machine models [19] when used on the same dataset. BERT has also been used to develop cyberbullying detection models using different datasets. In their study, Desai et al. [20] compared the performance of classical machine learning models to the BERT model using the Twitter Dataset from Dalvi et al. [21]. Their findings show significant accuracy differences where the Naive Bayes classifier and Support Vector Machine classifiers achieved 52.70% and 71.25% accuracy, respectively, and the BERT model had 91.90% accuracy [20].

Elsafoury et al. [22] contend that although BERT outperforms other models in cyberbullying detection, there needs to be more literature on the aspects contributing to BERT's outstanding performance. [22] used five datasets to determine the features that BERT relies on for its performance and the effect of attention weights on performance. The study examined classical machine learning models, fine-tuned BERT models, and non-fine-tuned BERT models and concluded that attention weights do not explain BERT's performance [22]. In this paper, the fine-tuned BERT model is trained on a large English dataset of tweets classified into seven categories; bullying, insult, spam, profanity, sarcasm, threat, pornographic content, and exclusion [23]. Multi-label classification enables more nuanced sentiment analysis since bullying on online platforms can take different forms.

Materials and Methods

This project will use a similar methodology to that used in standard machine learning tasks following the industry-wide practices adopted by researchers and organizations.

i) Data Collection

Data collection is vital because the data quality influences the machine learning model's performance. The project used a large dataset curated by [23] which consists of 62,587 tweets extracted from Twitter based on the likelihood of containing offensive content. Unlike some pre-existing datasets used in cyberbullying detection research, this dataset is multi-labelled and consists of seven categories; insult, profanity, sarcasm, threat, pornographic content, exclusion, and spam. In datasets such as the Twitter-Racism dataset, the tweets are labelled as either racist or not [22], while in the Twitter-Sexism dataset, the posts are classified as either sexist or not [22]. Binary classification datasets fail to consider broader contexts and have limited capabilities. Posts online can perpetuate bullying through different forms; thus, using multi-labelled data accounts for the likelihood that bullies can use text to intimidate victims in various ways. Multi-labelling facilitates the creation of a more effective cyberbullying detection system that can classify a text by considering a more comprehensive range of factors. Salawu et al. used 17 annotators from different racial backgrounds residing in different countries, and three annotators labelled each tweet to control for annotators' cultural and gender bias [23].

ii) Data Cleaning

Data cleaning involves pre-processing data to ensure it is complete and ready for use in subsequent processes for machine learning. Some data exploration was done to determine the data distribution and gain a summary of entries in the dataset. The dataset is imbalanced because it has few entries where the sentiments are labeled as threats, exclusionary and sarcastic. This could affect the model's ability to predict these classes.

iii) Model Selection

Model selection is essential when building and designing any machine learning system because different models have varying accuracies and are suitable for different tasks. This project will use the BERT base pre-trained model, which Nityasya et al. [24] identify as highly suitable for classification tasks.

iv) Model Training

The lack of sufficient training data is one of the most significant challenges in NLP because task-specific datasets are small, yet deep learning-based NLP models require more significant amounts of data for better performance [25]. Fine-tuning pre-trained models is one of the techniques used to address this challenge, where the model is pre-trained using enormous unannotated data and then fine-tuned for specific tasks [24]. A smaller dataset can be used in fine-tuning because the parameters learned from the pre-training are carried to the fine-tuning step; therefore, the model performs better than a model trained only on a specific task. BERT base(uncased) was fine-tuned by training for 10 epochs with a batch size of 10 and a learning rate of $2e{-}5$.

v) Model Evaluation

Model evaluation was done using various metrics to evaluate NLP systems, including accuracy, precision, recall and F1 score [26]. Accuracy is used in classification to determine the closeness of a measured value to the known value and where the output variable is categorical or discrete [26]. The precision determines the per cent of true positives identified given all instances the classifier labels positive. Recall evaluates how well a model can recall the positive class, and the recall value signifies the number of positive labels that the model has correctly identified as positive. The F1 score combines precision and recall into one metric, representing the model's performance.

Precision = true positive / (true positive + false positive)

Recall = true positive/ (true positive + false negative)

F1 Score = 2*((precision*recall)/ (precision + recall))

The Area Under Curve- Receiver Operating Characteristic (AUC-ROC) score shows the rank correlation between predictions and targets. It informs the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance. The ROC curve visualizes the differences between true and false positive rates [27].

Results

The initial training set for each model and dataset was randomly stratified-split into a training (70%) and test (30%) set. The model was then trained using the training set, validated and tested on the original test set. The performance of the fine-tuned BERT model is reported in Table 1.

Epoch	F1 Score	ROC AUC	Accuracy
1	0.817065	0.867197	0.659991
2	0.826612	0.882665	0.659459
3	0.821102	0.881296	0.657382
4	0.818248	0.881042	0.650085

The code used to achieve these results is shared in [28]

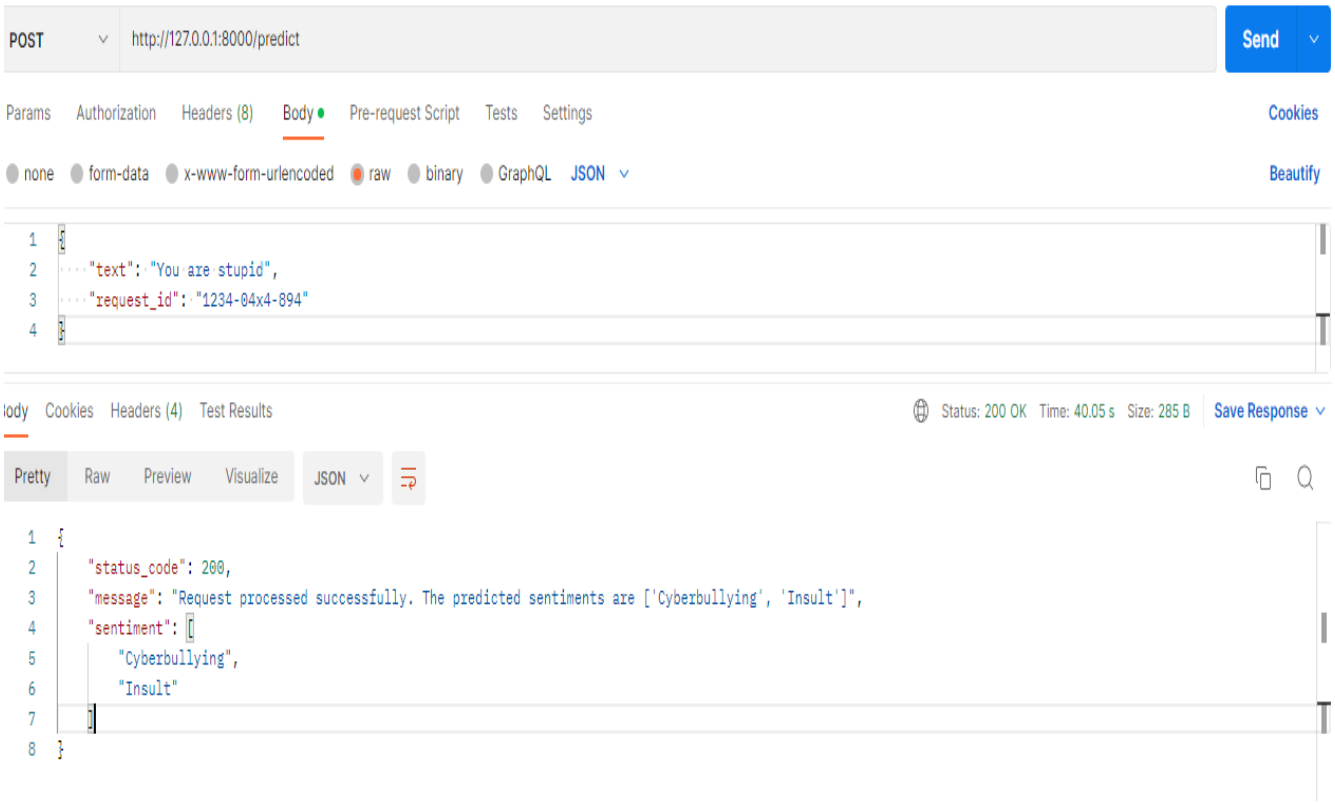

Figure 1.1: Request and response when a client requests the API.

The deployed version can be accessed in [29].

Discussion

The model achieves a 65% accuracy rate. The low accuracy can be attributed to the multi-label dataset, which requires the model to predict all seven labels for each entry accurately. However, accuracy is not very informative when working with multi-label data, especially when the dataset is imbalanced. The F1 score and the ROC-AUC provide better insights into the model's performance. The model achieved an F1 score of 81% and a ROC-AUC of 88%. This indicates that the model obtained a precision and recall rate leading to the high F1 score. Using a large dataset helped avoid techniques such as oversampling to provide the model with sufficient data to learn from and make proper sentiment classification.

The current model has some limitations that may affect its performance in production, especially when classifying posts with underrepresented sentiments in the training data. Noise in the dataset may also have affected the model's performance since the data is derived from Twitter which does not allow posts to exceed a certain number of characters. [22] highlight that the dataset's domain contributes to the amount of noise in the dataset, which affects the fine-tuned model's performance as it learns some syntactical biases. Additionally, perpetrators of cyberbullying often use images and text to target the victim, and this model cannot handle images that may provide better context on the sentiments expressed.

Conclusion

Cyberbullying has become widely prevalent with the advent of the internet and the widespread adoption of social media platforms. Perpetrators use different techniques to intimidate, humiliate or threaten victims making online spaces unsafe for users. Online bullying may involve threatening messages, offensive or aggressive texts, posting embarrassing images or videos of the victim or impersonating a person in an unflattering way. Incidences of bullying have been reported in Kenya, with famous people being the most susceptible to bullying, while ordinary people with a relatively low number of followers online also experience bullying. Children and adolescents are also susceptible to bullying because they use online platforms, and their peers and other general users can easily send hurtful or offensive posts or comments about them. Bullying can also occur in the workplace, affecting the employee's performance and ability to collaborate with colleagues. Cyberbullying is also linked to depression, anxiety and decreased performance and productivity.

The cyber-bullying detection system can predict whether posts are offensive or threatening by utilizing the BERT model, a state-of-the-art NLP model used for sentiment analysis. The service is designed to be a middle-tier service that other client services can utilize to improve user experience by minimizing incidences of bullying. This project contributes to making responsible software where the client services are mindful of their impact on society. Preventing cyberbullying is vital in increasing safety in cyberspace and maintaining spaces where everyone can express their opinions without infringing on others.

References

[1]"Africa Internet User Stats and 2022 Population by Country", Internetworldstats.com, 2022. [Online]. Available: https://www.internetworldstats.com/africa.htm#ke. [Accessed Jul 18, 2022].

[2] Kanyi, Patrick. (2020). The Kenyan Social Media Landscape. 10.13140/RG.2.2.11876.60809. [Accessed Jul 18, 2022].

[3] D. Otieno, F. Kirigha and A. Akwala, "Communication on Social Network Sites", Dialectical Perspectives on Media, Health, and Culture in Modern Africa, pp. 224-238, 2021. Available: 10.4018/978-1-5225-8091-1.ch013. [Accessed Jul 18, 2022].

[4]"Cyberbullying: What is it and how to stop it", Unicef.org, 2022. [Online]. Available: https://www.unicef.org/end-violence/how-to-stop-cyberbullying. [Accessed Jul 18, 2022].

[5] P. Ferrara, F. Ianniello, A. Villani and G. Corsello, "Cyberbullying a modern form of bullying: let's talk about this health and social problem", Italian Journal of Pediatrics, vol. 44, no. 1, 2018. Available: 10.1186/s13052-018-0446-4. [Accessed Jul 18, 2022].

[6] C. Zhu, S. Huang, R. Evans and W. Zhang, "Cyberbullying Among Adolescents and Children: A Comprehensive Review of the Global Situation, Risk Factors, and Preventive Measures", Frontiers in Public Health, vol. 9, 2021. Available: 10.3389/fpubh.2021.634909. [Accessed Jul 18, 2022].

[7] S. Shearman, "Cyberbullying in the workplace: 'I became paranoid'", the Guardian, 2017. [Online]. Available: https://www.theguardian.com/careers/2017/mar/30/cyberbullying-in-the-workplace-i-became-paranoid. [Accessed Jul 18, 2022].

[8] G. Abaido, "Cyberbullying on social media platforms among university students in the United Arab Emirates", International Journal of Adolescence and Youth, vol. 25, no. 1, pp. 407-420, 2019. Available: 10.1080/02673843.2019.1669059. [Accessed Jul 18, 2022].

[9] A. John et al., "Self-Harm, Suicidal Behaviours, and Cyberbullying in Children and Young People: Systematic Review", Journal of Medical Internet Research, vol. 20, no. 4, p. e129, 2018. Available: doi:10.2196/jmir.9044. [Accessed Jul 18, 2022].

[10] Z. Drus and H. Khalid, "Sentiment Analysis in Social Media and Its Application: Systematic Literature Review", Procedia Computer Science, vol. 161, pp. 707-714, 2019. Available: 10.1016/j.procs.2019.11.174. [Accessed Jul 18, 2022].

[11] I. El Alaoui, Y. Gahi, R. Messoussi, Y. Chaabi, A. Todoskoff and A. Kobi, "A novel adaptable approach for sentiment analysis on big social data", Journal of Big Data, vol. 5, no. 1, 2018. Available: 10.1186/s40537-018-0120-0. [Accessed Jul 18, 2022].

[12] M. Wankhade, A. Rao and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges", Artificial Intelligence Review, 2022. Available: 10.1007/s10462-022-10144-1 [Accessed Jul 18, 2022].

[13] S. Almutiry and M. Fattah, "Arabic Cyberbullying Detection Using Arabic Sentiment Analysis Arabic Cyberbullying Detection Using Arabic Sentiment Analysis", Egyptian Journal of Language Engineering, vol. 8, no. 1, 2021. Available: https://ejle.journals.ekb.eg/article_160441_43a97c5f5071efed9f9e25fdbb459180.pdf. [Accessed Nov. 1, 2022].

[14] H. Rosa et al., "Automatic cyberbullying detection: A systematic review", Computers in Human Behavior, vol. 93, pp. 333-345, 2019. Available: 10.1016/j.chb.2018.12.021 [Accessed Nov. 1, 2022].

[15] J. Devlin, M. Chang, K. Kenton Lee and K. Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", Google AI Language, 2019. Available: http://arxiv.org/pdf/1810.04805.pdf. [Accessed Nov. 1, 2022].

[16] A. Gillioz, J. Casas, E. Mugellini and O. Khaled, "Overview of the Transformer-based Models for NLP Tasks", Proceedings of the 2020 Federated Conference on Computer Science and Information Systems, 2020. Available: 10.15439/2020f20. [Accessed Nov 1, 2022].

[18] A. Vaswani, N. Shazier, N. Niki Parmar and J. Uszkoreit, "Attention Is All You Need", 2017. Available: https://arxiv.org/pdf/1706.03762.pdf. [Accessed Nov. 1, 2022].

[19] M. Bilal and A. A. Almazroi, "Effectiveness of fine-tuned BERT model in the classification of helpful and unhelpful online customer reviews," Electronic Commerce Research, 2022. [Accessed Nov. 1, 2022].

[20] A. Desai, S. Kalaskar, O. Kumbhar, and R. Dhumal, "Cyberbullying detection on social media using machine learning," ITM Web of Conferences, vol. 40, Aug. 2021. [Accessed Nov. 1, 2022].

[21] R. R. Dalvi, S. Baliram Chavan, and A. Halbe, “Detecting a Twitter cyberbullying using machine learning,” 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), May 2020.

[22] F. Elsafoury, S. Katsigiannis, S. R. Wilson, and N. Ramzan, “Does Bert Pay Attention to Cyberbullying?” Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021. [Accessed Nov. 1, 2022]

[23] S. Salawu, J. Lumsden, and Y. He, “A large-scale English multi-label Twitter dataset for cyberbullying and online abuse detection,” Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021), 2021. [Accessed Nov. 1, 2022]

[24] M. N. Nityasya, H. A. Wibowo, R. Chevi, R. E. Prasojo, and A. F. Aji, “Which Student is Best? A Comprehensive Knowledge Distillation Exam for Task-Specific BERT Models.,” Jan. 2022. [Accessed Nov. 1, 2022]

[25] S. Khalid, "Bert explained: A Complete Guide with Theory and tutorial," Medium, 10-Apr-2020. [Online]. Available: HTTPS://medium.com/@samia.khalid/bert-explained-a-complete-guide-with-theory-and-tutorial-3ac9ebc8fa7c. [Accessed Nov. 1, 2022]

[26] S. Yashaswini and S. S. Shylaja, “Metrics for automatic evaluation of text from NLP models for text to scene generation,” European Journal of Electrical Engineering and Computer Science, vol. 5, no. 4, pp. 20–25, 2021. [Accessed Nov. 1, 2022]

[27] “Classification: Roc curve and AUC,” Google, 18-Jul-2022. [Online]. Available: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc. [Accessed Nov. 1, 2022].

[28] B. Waweru, “Cyberbullying detection” Github. Available: https://github.com/wanguiwaweru/Cyberbullying-detection [Accessed Nov. 17, 2022].

[29] B. Waweru, “Cyberbullying detection”. Available: http://cyberbullying.eastus2.cloudapp.azure.com:8000/docs [Accessed Nov. 17, 2022].