Detecção de Cyberbullying em Redes Sociais

João Vitor Ferreira[1], **Helen C. de Mattos Senefonte**[1]

[1]Departamento de Computação – Universidade Estadual de Londrina (UEL)
Caixa Postal 10.011 – CEP 86057-970 – Londrina – PR – Brasil

abc-joao@hotmail.com, helen@uel.br

Abstract. *With the increase in the use of social networks, new research opportunities are emerging, people expose their personal information, their contacts and also facts and information about their lives. Despite the positive side of maintaining these virtual relationships, the exposure of these personal data also brings negative points, such as the use of them to invade their owners' personal accounts, locate people with malicious purposes, use phone numbers for false kidnappings, haters, who denigrate the image of many people shamelessly. Early detection and identification of such events can counter the threat of this unethical practice. The objective of this work is to propose a model capable of detecting these threats and helping the cyberbully prevention process.*

Resumo. *Com o aumento do uso das redes sociais, novas oportunidades de pesquisa vêm surgindo, as pessoas expõem suas informações pessoais, seus contatos e também fatos e informações sobre suas vidas. Apesar do lado positivo de se manter essas relações virtuais, a exposição desses dados pessoais traz também pontos negativos, como por exemplo a utilização deles para invadir contas pessoais de seus proprietários, localizar pessoas com fins maliciosos, utilizar números telefônicos para falsos sequestros, ataque de haters*[1]*, os qual denigrem a imagem de muitas pessoas sem pudor. A detecção precoce e a identificação de tais eventos podem conter a ameaça dessa prática antiética. O objetivo desse trabalho é propor um model capaz de detectar essas ameaças e auxiliar o processo de prevenção do cyberbully.*

1. Introdução

Atualmente enfrentamos um grande problema social: o bullying. Com o avanço da tecnologia e com sua repercussão em várias áreas do cotidiano das pessoas, o bullying não se limita apenas ao mundo real, mas se estende para o mundo virtual, conhecido como cyberbullying. Além das agressões sofridas presencialmente, várias pessoas sofrem assédio em redes sociais, seja por meio de mensagens, fotos e até videos ameaçando-as. O cyberbullying [19] pode surgir de várias motivações diferentes, como aparência física, racismo, sexismo, inteligência, entre outros.

Com o aumento do uso de plataformas digitais para fins de educação, lazer, político, econômico, ocorre também o aumento nos incidentes de cyberbullying. Parece inevitável, pois os alunos que sofrem bullying estão mais propensos ao cyberbully, então ocorre, paralelamente, um uso cada vez maior no uso de redes sociais [17]. O bullying e a discriminação com base em raça, religião, sexo, casta e credo podem causar um efeito

[1]Termo inglês (em tradução direta significa: "Odiadores") utilizado para classificar aqueles que praticam bullying virtual.

adverso na saúde mental da vítima, levando eventualmente à ansiedade, depressão e até mesmo ao aumento dos casos de suicídio.

Detectar o cyberbullying e identificar quem o fez não é uma tarefa simples e na literatura atual, existem algumas técnicas e conceitos para executar essa árdua tarefa. Em geral são utilizados modelos de Machine Learning (ML), tais como Naive Bayes e Support Vector Machine (SVM), além de modelos de Processamento de Linguagem Natural (PLN) como o BERT (Bidirectional Encoder Representations from Transformers).

Através de um grande volume de comentários capturados em redes sociais, este trabalho tem como objetivo identificar e classificar comentários considerados bullying, utilizando técnicas e conceitos de Machine Learning, a fim de classificá-los de forma acertiva e definir um nível de bullying cometido.

Este documento está organizado da seguinte forma: a Seção 2 apresenta os principais conceitos necessários para a compreensão do projeto. Na seção 2.1 é realizado um levantamento bibliográfico preliminar de trabalhos correlatos A Seção 3 apresenta os objetivos do projeto. As principais atividades propostas são descritas na Seção 4. Na seção 5 é proposto um planejamento preliminar das atividades a serem desenvolvidas. A seção 6 expõe uma visão do que se espera de resultados deste trabalho e em que ele pode agregar de conhecimento sobre o assunto.

2. Fundamentação Teórico-Metodológica e Estado da Arte

Algumas definições são apresentadas nessa seção, visando o melhor entendimento dos modelos que serão explorados no projeto. Inicialmente o conceito de machine learnning, termo comumente utilizado *Machine Learning (ML)*.

Tom M. Michell (1997, p. 3) [10] define machine learning como: “Campo de estudo que dá aos computadores a capacidade de aprender sem serem explicitamente programados“. Um exemplo disso é um computador que aprende a jogar xadrez melhorando seu desempenho a partir das vitórias em cada partida.

Aurélio Gerón(2019, p. 4) [7] define que “Machine Lerarning (ML) é a ciência (e arte) de programar computadores para que eles possam aprender com dados“. Já Arthur Samuel (1959) afirma que “[Machine Learning é o] campo de estudo que dá aos computadores a capacidade de aprender sem serem explicitamente programados“. Para que o computador aprenda deve haver uma inspeção no conjunto de dados, uma identificação de padrões, para que o computador possa tormar decisões sem tanta programação.

No mundo do ML, outro termo muito utilizado é *Data Mining*, na qual, neste contexto seria a aplicação de técnicas de ML para explorar grandes quantidades de dados podendo ajudar a descobrir padrões que não eram imediatamente aparentes [7].

Outro conceito importante para o desenvolvimento desse projeto é o de *Artificial Neural Networks*. Segundo Michell [10], “(...) o estudo das redes neurais artificiais (RNAs) foi inspirado em parte pela observação de que os sistemas de aprendizado biológico são construídos de redes muito complexas de neurônios interconectados. Em analogia grosseira, as redes neurais artificiais são construídas a partir de um conjunto densamente interconectado de unidades simples, onde cada unidade recebe um número de entradas de valor real (possivelmente as saídas de outras unidades) e produz uma única

saída de valor real (que pode se tornar a entrada para muitas outras unidades)" (Tom M. Michell, 1997).

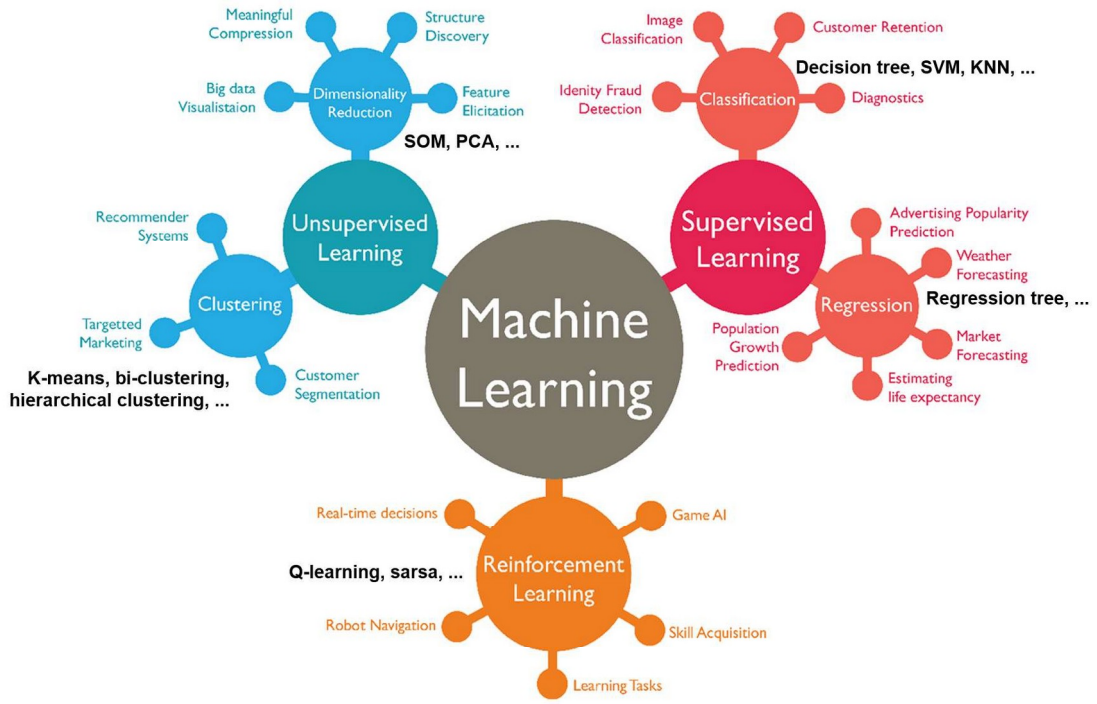

Figura 1. Diagrama que apresenta os diferentes tipos de Machine Learning

Na figura 1, machine learning (o circulo maior) está subdividida em 3 categorias (os 3 circulos um pouco menores), de acordo com a quantidade e o tipo de supervisão que recebe [7], que são: *supervised learning, unsupervised learning* e *reinforcement rearning*. Supervised learning você treina o algoritmo com dados factíveis sobre, por exemplo, uma bicicleta, e ele mapeia uma função de entrada e saída [7]. Ou seja, são definidos atributos e seus valores sobre essa bicicleta (bicicleta possui um quadro que mede 60 cm, rodas de aro 18, pedais de ferro, ...). Um clássico exemplo de supervised learning é o filtro de spam dos emails. Ele é treinado com muitos emails, que você pode classifica-los como spam. Unsupervised learning o algoritmo define padrões da entrada de dados, ainda que não haja nenhuma informação fornecida previamente, isto é, o algoritmo tenta aprender sem que tenha quem o ensine[18]. Nesta categoria temos uma representação mais informativa e simples dos dados, condensando a informação em pontos mais relevantes. Um exemplo prático hoje, ocorrem no site Linkedin [2], na qual ele surege conexões para você a partir de dados do usuário. No reinforcement learning, existe um sistema de aprendizagem chamado de agente, na qual ele seleciona uma ação a ser executada e receber uma recompença ou punição em troca. Logo, o agente aprende qual a melhor ação a se tomar e cria uma política de quais ações tomar em cada situação [7]. Um exemplo disso são as recomendações de videos do Youtube [3], onde ele te sugere um vídeo e, se você o assiti-lô todo, o agente aprende que o video é relevante para você (recompensa), porém, caso você, comece o video e logo troque de video, o agente assume que você não gostou do video (punição) e não o recomenda mais.

[2]Rede social para profissionais que estão a procura de emprego

[3]Plataforma de vidoes online.

A figura 1 também ilustra os tipos de machine learning e os modelos mais populares de cada abordagem. Alguns dos modelos são brevemente descritos a seguir:

- **PCA**: *Principal Component Analysis (PCA)* é um algoritmo estatístico usado para transformar um conjunto de variáveis possivelmente correlacionadas em um conjunto de recombinações lineares não correlacionadas dessas variáveis chamadas de componentes principais[7].
- **SOM**: *Self Organizing Maps (SOM)* é uma técnica de dimensionamento multidimensional que constrói uma aproximação da função de densidade de probabilidade de algum conjunto de dados subjacente, que também preserva a estrutura topológica desse conjunto de dados.
- **Decision Tree**: é uma estrutura de árvore semelhante a um fluxograma, onde cada nó interno denota um teste em um atributo, cada ramo representa um resultado do teste e cada nó folha (nó terminal) contém um rótulo de classe [7].
- **SVM**: *Support Vector Machine* é um algoritmo de machine learning que analisa dados para classificação e análise de regressão. O SVM é um método de aprendizado supervisionado que analisa os dados e os classifica Um SVM gera um mapa dos dados classificados com as margens entre os dois o mais distantes possível. SVMs são usados na categorização de texto, classificação de imagens, reconhecimento de escrita e nas ciências [18].
- **KNN**: *K-Nearest Neighbors* é um dos muitos algoritmos (de aprendizagem supervisionada) usado no campo de data mining e machine learning, ele é um classificador onde o aprendizado é baseado "no quão similar" é um dado (um vetor) do outro. O treinamento é formado por vetores de n dimensões [7].
- **K-means**: É um algoritmo de aprendizado não supervisionado que avalia e clusteriza os dados de acordo com suas características. Na prática ele tenta separar os dados em k clusters, os dados geralmente têm que estar na forma de vetores numéricos. Estritamente falando, o método funcionará desde que você tenha uma maneira de calcular a média de um conjunto de pontos de dados e a distância euclidiana entre eles [7].
- **Bi-clustering**: são tarefas de mineração de dados capazes de extrair informações relevantes dos dados aplicando critérios de similaridade simultaneamente a linhas e colunas de matrizes de dados. Algoritmos utilizados para realizar essas tarefas agrupam simultaneamente objetos e atributos, possibilitando a descoberta de biclusters[7].
- **Hierarchical Clustering**: é um algoritmo que agrupa objetos semelhantes em grupos chamados clusters . O endpoint é um conjunto de clusters , em que cada cluster é distinto do outro cluster e os objetos dentro de cada cluster são amplamente semelhantes entre si [7].
- **regression tree**: é construída por meio de um processo conhecido como particionamento recursivo binário, que é um processo iterativo que divide os dados em partições ou ramificações e continua dividindo cada partição em grupos menores à medida que o método avança em cada ramificação [7].
- **Q-learning**: uma política de aprendizado de reforço que encontrará a próxima melhor ação, dado um estado atual. Ele escolhe essa ação aleatoriamente e visa maximizar a recompensa. É um aprendizado de reforço fora da política, sem modelo, que encontrará o melhor curso de ação, dado o estado atual do agente.

Dependendo de onde o agente estiver no ambiente, ele decidirá a próxima ação a ser tomada [18].

- **Sarsa**: O algoritmo SARSA *State–action–reward–state–action* é uma pequena variação do popular algoritmo Q-Learning e é um dos algoritmos de aprendizado por reforço que aprende com o conjunto atual de estados e ações e aprende com a mesma política de destino. O principal ponto que diferencia o algoritmo SARSA do algoritmo Q-learning é que ele não maximiza a recompensa para o próximo estágio de ação a ser realizado e atualiza o valor Q para os estados correspondentes [18].

2.1. Trabalhos Correlatos

Nessa seção, é apresentado o levantamento bibliográfico realizado dentre os trabalhos correlatos, visando direcionar as investigações desse projeto.

Em [5], os autores apresentam uma abordagem para detecção não supervisionada de cyberbullying. O modelo proposto consiste em dois componentes principais: uma *Representation Learning Network* que codifica a sessão de mídia social explorando recursos multimodais, por exemplo, texto, rede e tempo; e a *Multi-task Learning Network* que ajusta simultaneamente os tempos entre chegadas de comentários e estima a probabilidade de bullying com base em um modelo de mistura gaussiana. Foi utilizado uma estrutura de detecção de cyberbullying não supervisionada chamada UCD (Unsupervised Cyberbullying Detection) via Time-Informed Gaussian Mixture Model, resultados experimentais em dois conjuntos de dados do mundo real corroborar a eficácia da UCD.

Em [2, 1] é abordado o problema da análise de sensibilidade de conteúdo direto. Através de deep learning e sequential deep neural network models, vários posts textuais são analisados e o usuário é classificado como sendo ou não cyberbully.

Em [16] foi proposto a primeira abordagem multitarefa que aproveita o conhecimento afetivo compartilhado para detectar discurso de ódio em tweets em espanhol, usando transformer-based model. Os resultados mostram que a combinação de polaridade e conhecimento emocional ajudam a detectar discursos de ódio com mais precisão nos conjuntos de dados.

Em [6], os autores focam na utilização de modelos de detecção de HS (Hate Speech) para aplicar na linguagem Roman Urdu, através de dicionario de palavras, mapeamento de gírias e tokenização [4]. A detecção foi realizada através da implementação de modelos de RNN-LSTM (*Recurrent Neural Networks - Long Shot Term Memory*), RNN-BiLSTM (*Recurrent Neural Networks - Bidirectional Long Shot Term Memory*) e CNN (*Convolutional Neural Network*).

Em [15] os autores propoem um sistema de detecção de cyberbullying multilíngue para detecção de cyberbullying em duas línguas indianas: Hindi e Marathi. O artigo propõe uma solução utilizando LR (*Logistic Regression*) a qual, funciona bem para classificação de usuários e melhora a medida que o tamanho dos dados de entrada aumentam. No artigo também é utilizado SGD (*Stochastics Gradient Descent*), na qual executa mais rápido que a solução usando LR, porém em LR, quanto maior a base de dados, menor é o erro.

[4]O processo de tokenização gera tokens de correspondência que são usados subsequentemente pelo processo de correspondência para identificar os registros de objeto base candidatos para correspondência.

Em [11] o artigo apresenta uma abordagem para detectar cyberbullying em streams árabes do Twitter. A metodologia proposta lê as mensagens do twitter em tempo real, processa e limpa os tweets do ruído, detecta mensagens ofensivas e as classifica de acordo com sua força atribuindo pesos a cada mensagem de bullying.

Em [9] o trabalho tem como objetivo propor um léxico de cyberbullying para mídias sociais e se concentra no cyberbullying de exclusão e propõe um léxico de cyberbullying de exclusão usando uma abordagem ontológica. O léxico proposto pode ser usado como um dicionário para os usuários nas mídias sociais.

Em [8] o artigo descreve as técnicas e recursos utilizadas na detecção do cyberbullying. São revisadas as fontes de dados disponíveis, os recursos e as técnicas de classificação utilizadas. NLP (*Natural Language Processing*) e *Machine Learnnig* são as famosas abordagens usadas para identificar palavras-chave de bullying dentro do corpus.

Em [20] o artigo se concentra na detecção de cyberbullying em páginas com termos em espanhol e utiliza técnicas de classificação de sentimento, conjunto de termos pejorativos, na análise são utilizadas técnicas de mineração de dados para gerar um dicionário. No artigo são mencionadas algumas plataformas (como por exemplo o Mr. Tweet e o Sentimento140) para essas etapas e classifica quais são mais flexíveis para otimizar o trabalho de detecção. Foi concluído que para detectar o cyberbullying é necessário analisar o contexto e os padrões semânticos de cada frase.

Em [12] é construído um modelo de classificação com ótima precisão na identificação de conversas de cyberbullying usando o método *Naive Bayes* e SVM (*Support Vector Machine*). Utilizando n-gram de 1 a 5 para o número de classes 2, 4, 11 Naive Bayes produz uma precisão média de 92,81%, SVM com um poly kernel produz uma precisão média de 97,11%.

Em [14] o artigo concentra-se na detecção de cyberbullying em linguagens com troca de código, como acontece na Índia, por exemplo,na qual tem-se uma sociedade multilingue. São utilizados diferentes algoritmos de machine learning, como (*Support Vector Machine*) e (*Logistic Regression*) e *Deep Learnnig* (*Multilayer Perceptron, Convolution Neural Network, BiLSTM, BERT*) para detectar cyberbullying de inglês-hindi (En-Hi) texto comutado por código.O modelo proposto tem resultado de 0,93 na pontuação F1com média macro.

Em [17] o trabalho propõe uma estrutura de *deep learning* que avaliará tweets ou postagens de mídia social em tempo real, bem como identificará corretamente qualquer conteúdo de cyberbullying neles. No artigo diz que em estudos recentes mostraram que as abordagens baseadas em redes neurais profundas são mais eficazes do que as técnicas convencionais na detecção de textos de cyberbullying. CNN sozinha só pode treinar características locais a partir de n-grams de palavras, com sua camada LSTM, a CNN-BiLSTM também pode aprender recursos globais e dependências de longo prazo. CNN (*Convolutional Neural network*) sozinha só pode treinar características locais a partir de n-grams de palavras, com sua camada LSTM (*Long short-term memory*), a CNN-BiLSTM (*Bidirectional Long Short-term Memory*) também pode aprender recursos globais e dependências de longo prazo. Os resultados mostrados utilizando modelos de redes neurais (CNN-BiLSTM) tem a melhor precisão.

Em [3, 5, 4] a discussão é centrada no desafio de modelar padrões temporais de

comportamento de cyberbullying. Investiga como a informação temporal dentro de uma sessão de mídia social, que tem uma estrutura inerentemente hierárquica (por exemplo, palavras formam um comentário e comentários formam uma sessão), podem ser aproveitadas para facilitar a detecção de cyberbullying.

Em [13] foi apresentado uma nova aplicação do BERT (Bidirectional Encoder Representations from Transformers) para identificação de cyberbullying. Um modelo de classificação simples usando BERT é capaz de alcançar resultados de última geração em três corpora do mundo real: Formspring, Twitter e Wikipedia. Os resultados experimentais demonstram que nosso modelo utilizando BERT teveos seguintes resultados na pontuação F1: 0.96, 0.94, 0.94; para os seguintes consjuntos de dados: Twitter, Wikipédia e FormSpring respectivamente. Ele alcança melhorias significativas em relação aos trabalhos existentes, em comparação com os modelos de redes neurais profundas baseados em slots ou atenção.

Em [21] são usadas CNN's (*Convolutional Neural Network*) personalizadas para o processamento de comentários de usuários do Instagram atraves de um sistema web. Os resultados mostram um acerto na detecção de 84,29% para cyberbullying e 83,08% para cyberagressão.

3. Objetivos

Este trabalho de pesquisa visa desenvolver um modelo que possa detectar com precisão o cyberbully em redes sociais. Os objetivos específicos são descritos a seguir:

- Extrair padrões de comportamento da base de dados, buscando melhorar e facilitar a detecção de comportamentos bullies;
- Definir um modelo analítico que julgue corretamente as situações que ocorreram o bullying;
- Sinalizar ao sistema que houve cyberbullying para que o mesmo aplique suas ações e respectivas sanções.

4. Procedimentos metodológicos/Métodos e técnicas

Para alcançar os objetivos propostos nesse projeto, serão elaboradas as atividades descritas a seguir:

- **Revisão da literatura**: Nesta etapa do trabalho será feito um levantamento da literatura correlata.
- **Modelagem da solução**: A partir da analise da literatura apresentada será proposto uma solução para a detecção de cyberbulling em mídias socias. Esta solução terá como base algum(ns) modelo(s) já proposto(s) na literatura, com a aquisição dos conhecimentos observados aqui.
- **Obtenção dos dados para os datasets**: Nesta etapa será feita a aquisição dos dados usados no treinamento dos modelos de ML a serem implementados.
- **Implementação do modelo**: Dado o modelo proposto no item anterior, será implemetado um algoritmo na linguagem Python 3 [5] que solucione o problema deste trabalho.
- **Análise e validação dos resultados**: Buscando um resultado satisfatório para o trabalho, os dados serão validados e analisados.

[5]Python 3 é uma linguagem de programação com vários recursos necessários para implementação de modelos com IA.

5. Cronograma de Execução

Nesta etapa é mostrada, uma macrovisão de quais as atividades serão executados no decorrer deste trabalho, através da tabela abaixo.

Atividades:

1. Revisão da literatura;
2. Modelagem da solução;
3. Obtenção dos dados para os datasets;
4. Implementação do modelo;
5. Análise e validação dos resultados;

Tabela 1. Cronograma de Execução

	set/22	out/22	nov/22	des/22	jan/23	fev/23	mar/23	abr/23	mai/23
Atividade 1	X	X							
Atividade 2		X	X						
Atividade 3			X	X	X				
Atividade 4				X	X	X	X		
Atividade 5							X	X	X

6. Contribuições e/ou Resultados esperados

Com o trabalho finalizado, espera-se que a solução feita possa ajudar a comunidade acadêmica na prevenção, identificação e classificação de cyberbullying em mídias sociais. Espera-se que a implementação possa identificar se houve algum tipo de violência em diálogos, posts ou comentários na redes sociais.

7. Espaço para assinaturas

Londrina, *12/09/2022*.

João Victor Ferreira

Aluno

Orientador

Referências

[1] Mohamed Berrimi, Abdelouaheb Moussaoui, Mourad Oussalah, and Mohamed Saidi. Attention-based networks for analyzing inappropriate speech in arabic text. In *2020 4th International Symposium on Informatics and its Applications (ISIA)*, pages 1–6, 2020.

[2] Livio Bioglio and Ruggero G. Pensa. Analysis and classification of privacy-sensitive content in social media posts. *EPJ Data Science*, 2021.

[3] Lu Cheng, Ruocheng Guo, Yasin Silva, Deborah Hall, and Huan Liu. *Hierarchical Attention Networks for Cyberbullying Detection on the Instagram Social Network*, pages 235–243. Society for Industrial and Applied Mathematics, 2019.

[4] Lu Cheng, Ruocheng Guo, Yasin N. Silva, Deborah Hall, and Huan Liu. Modeling temporal patterns of cyberbullying detection with hierarchical attention networks. *Society for Industrial and Applied Mathematics*, 2(2), 2021.

[5] Lu Cheng, Kai Shu, Siqi Wu, Yasin N. Silva, Deborah L. Hall, and Huan Liu. Unsupervised cyberbullying detection via time-informed gaussian mixture model. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 185–194, New York, NY, USA, 2020. Association for Computing Machinery.

[6] Amirita Dewani, Mohsin Ali Memon, and Sania Bhatti. Cyberbullying detection: advanced preprocessing techniques & deep learning architecture for roman urdu data. In *2020 4th International Symposium on Informatics and its Applications (ISIA)*. Journal of Big Data, 2020.

[7] Aurélien Géron. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*, volume 3. McGraw-Hill Science/Engineering/Math, 2019.

[8] Wan Noor Hamiza Wan Ali, Masnizah Mohd, and Fariza Fauzi. Cyberbullying detection: An overview. In *2018 Cyber Resilience Conference (CRC)*, pages 1–3, 2018.

[9] Ong Chee Hang and Halina Mohamed Dahlan. Cyberbullying lexicon for social media. In *2019 6th International Conference on Research and Innovation in Information Systems (ICRIIS)*, pages 1–6, 2019.

[10] Tom M. Mitchell. *Machine Learning*, volume 1. McGraw-Hill Science/Engineering/Math, 1997.

[11] Djedjiga Mouheb, Masa Hilal Abushamleh, Maya Hilal Abushamleh, Zaher Al Aghbari, and Ibrahim Kamel. Real-time detection of cyberbullying in arabic twitter streams. In *2019 10th IFIP International Conference on New Technologies, Mobility and Security (NTMS)*, pages 1–5, 2019.

[12] Noviantho, Sani Muhamad Isa, and Livia Ashianti. Cyberbullying classification using text mining. In *2017 1st International Conference on Informatics and Computational Sciences (ICICoS)*, pages 241–246, 2017.

[13] Sayanta Paul and Sriparna Saha. Cyberbert: Bert for cyberbullying identification. *Springer Nature*, 2020.

[14] Sayanta Paul, Sriparna Saha, and Jyoti Prakash Singh. Covid-19 and cyberbullying: deep ensemble model to identify cyberbullying from code-switched languages during the pandemic. *Springer Science*, 2021.

[15] Rohit Pawar and Rajeev R. Raje. Multilingual cyberbullying detection system. In *2019 IEEE International Conference on Electro Information Technology (EIT)*, pages 040–044, 2019.

[16] Flor Miriam Plaza-Del-Arco, M. Dolores Molina-González, L. Alfonso Ureña-López, and María Teresa Martín-Valdivia. A multi-task learning approach to hate speech detection leveraging sentiment analysis. *IEEE Access*, 9:112478–112489, 2021.

[17] Mitushi Raj, Samridhi Singh, Kanishka Solanki, and Ramani Selvanambi. An application to detect cyberbullying using machine learning and deep learning techniques. *EPJ Data Science*, 2022.

[18] Stuart Russel and Peter Norvig. *Artificail Intelligence, A modern Approach*, volume 2. Pearson, 2020.

[19] Peter K. Smith, Jess Mahdavi, Manuel Carvalho, Sonja Fisher, Shanette Russell, and Neil Tippett. Cyberbullying: its nature and impact in secondary school pupils. *Journal of Child Psychology and Psychiatry*, 49(4):376–385, 2008.

[20] Freddy Tapia, Cristina Aguinaga, and Roger Luje. Detection of behavior patterns through social networks like twitter, using data mining techniques as a method to detect cyberbullying. In *2018 7th International Conference On Software Process Improvement (CIMPS)*, pages 111–118, 2018.

[21] Haoti Zhong, David J Miller, and Anna Squicciarini. Flexible inference for cyberbully incident detection. In Ulf Brefeld, Edward Curry, Elizabeth Daly, Brian MacNamee, Alice Marascu, Fabio Pinelli, Michele Berlingerio, and Neil Hurley, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 356–371, Cham, 2019. Springer International Publishing.