Plan van Aanpak - Blok 2

GOGCAY KAAN

ACM Reference Format:
Gogcay Kaan. 2025. Plan van Aanpak - Blok 2. 1, 1 (December 2025), 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Context

In 2021 is er research gedaan naar online toxicity binnen gaming communities. Zo blijkt dat online toxicity een groot probleem is [2]. Toxicity kan worden gedefinieerd als het uiten van negatieve opmerkingen tegenover anderen. Maar niet alleen binnen gaming communities is toxicity een probleem. Sociale media zit helaas, net als de gaming communities, bomvol met haat, intimidatie en andere vormen van toxicity [18].

Maar waarom is toxicity zo een groot probleem? Toxicity is te herkennen in vele verschillende vormen. Vrijwel de meest problematische sub categorie van online toxicity is cyber-pesten, oftewel online pesten. Nu blijkt uit onderzoek dat kinderen die gepest zijn problemen ervaren omtrent gezondheid, emotionele welzijn en school prestaties. Gepeste kinderen voelen zich eerder angstig [4, 7, 11], depressief [4, 11, 16, 21] en hebben eerder een laag zelfbeeld [6, 9, 16, 21]. Helaas zijn de negatieve gevolgen van pesten niet gelimiteerd tot mentale klachten. Uit onderzoek blijkt dat pesten ook de oorzaak is voor vele fysieke klachten, zoals: buikpijn, slaap problemen, hoofdpijn, gespannenheid, bedplassen, uitputting, weinig eetlust [7].

Maar waarom zijn mensen toxic? John Suler beweert dat er zes factoren zijn waardoor mensen online geneigd zijn zo gezegd "de controle te verliezen [19]". Redenen hiervoor zijn:

- **Anonimiteit** - Je werkelijke identiteit is verborgen waardoor mensen zich minder snel verantwoordelijk voelen voor hun acties.
- **Onzichtbaarheid** - Je ziet elkaar niet fysiek waardoor sociale remmingen verminderen.
- **Asynchrooniteit** - Communicatie gebeurt niet in het moment waardoor je minder rekening houdt met de gevolgen.
- **Solipsistische introjectie** - Je projecteert je eigen gedachten en gevoelens op anderen.
- **Dissociatieve verbeelding** - Je voelt je losgekoppeld van het echte leven, alsof je je bevind in een fantasie.
- **Minimalisering van status en autoriteit** - Je status in de echte wereld is niet herkenbaar in de online wereld.

2 Doel en Oplossing

Mijn uiteindelijke doel is om een model te bouwen dat toxicity kan detecteren om zo toxicity online te verminderen. Een mogelijke aanpak om dit te bereiken is om het model op te splitsen in een taal classificatie model, en een model dat daadwerkelijk teksten kan classificeren op toxicity. Met deze aanpak kan het model eerst een inschatting maken in

Author's Contact Information: Gogcay Kaan.

Manuscript submitted to ACM

welke taal de toxic opmerking is gemaakt en vervolgens kan de geclassificeerde taal meegenomen worden als parameter in de toxicity classifier, om zo de toxicity classifier meer context te geven bij het inschatten van toxicity. Deze splitsing is een veel voorkomend fenomeen en wordt ook gebruikt in andere toepassingen [5, 13, 15].

Ik heb ervoor gekozen om enkel één van de twee modellen uit te werken. Dit is zo gekozen om de scope binnen dit research zoveel mogelijk af te bakenen, én is het onrealistisch om twee modellen volledig uit te werken in de gegeven tijd. Hieruit vormt mijn researchvraag: "Hoe ontwikkel je een taalmodel dat talen kan classificeren?"

3 Target Audience

Mijn doelgroep heeft geen duidelijke leeftijd, ras, culturele achtergrond, gender of andere voordehand liggende factor. De primaire doelgroep van mijn onderzoek bestaat uit gebruikers die actief deelnemen aan online platforms waar communicatie mogelijk is, zoals: sociale media, online gaming, forums en andere digitale omgevingen waarin tekstuele communicatie mogelijk is.

Deze doelgroep wordt op verschillende manieren benadeeld afhankelijk van de sub-doelgroep. Binnen gaming communities uit toxicity zich vaak in de vorm van communicatie, waaronder: intimidatie, verbaal geweld, en beledigingen, maar ook storende gameplay, zoals: griefing, spammen, en valsspelen [1, 8, 12, 17, 20]. Binnen sociale media kan toxicity voorkomen in de vormen van cyber-pesten, haatzaaien, desinformatie, complotten, extremisme, intimidatie, geweld [18].

Het doel voor de doelgroep is om hen te helpen voor een veiligere online ervaring. Dat kan worden gerealiseerd door dit project in te zetten om tekst met toxicity weg te filteren.

4 Topic Scope

De scope binnen dit project is het classificeren van een select aantal talen. De gekozen talen zijn gekozen op basis van welk alfabet de taal gebruikt. Zo neem ik alleen talen mee die gebruik maken van het Latijnse alfabet. Ook belangrijk om aan te kaarten is dat de classificatie van toxic berichten buiten dit project valt. Binnen dit project ligt de focus enkel op het classificeren welke taal een tekst is.

Mijn onderwerp en scope zijn indirect gelinkt aan het huidige groepsproject in blok 2. Zo wil ik classificeren welke taal er wordt gebruikt in een tekst om zo vervolgens de tekst te detecteren op toxicity. Dit sluit aan op het groepsproject omdat we binnen het groepsproject niet toxicity moeten detecteren maar B1 taal [14]. Zo is er in beide casussen een bepaalde stuk natuurlijke taal dat gedetecteerd moet worden.

5 Background Information

Taal classificatie wordt niet enkel gebruikt voor het detecteren van toxicity. Marco Lui bespreekt in zijn paper hoe taal identificatie, een synoniem voor taal classificatie, wordt gebruikt in het proces van personalisatie binnen sociale media. Bijvoorbeeld dat je voornamelijk comments leest onder posts in het Nederlands of in het Engels. Ook wordt deze technologie gebruikt voor het automatisch vertalen van tekst en ook voor het filteren van websites die content bevatten die je niet kunt lezen op basis van je zoekopdracht [13]. Bijvoorbeeld: met een Engelse zoekopdracht krijg je niet gauw een website met Russische content.

In ander onderzoek worden de vele belangrijke toepassingen van tekst classificatie ook benoemd, bijvoorbeeld: web search, information retrieval, ranking en document classificatie [5, 15]. Zo blijkt dus dat tekst classificatie een essentieel onderdeel is voor vele geavanceerde toepassingen.

6 Methodologie

Voor het trainen van een model dat tekst kan classificeren ga ik de volgende stappen ondernemen.

6.1 Data Analyse

Ik begin bij het zoeken van een dataset. Ik heb tijdens het schrijven van mijn plan van aanpak al een dataset gevonden die 32 miljoen comments bevat van 20 miljoen unieke users in 50 verschillende talen. Deze comments komen van 178,000 YouTube video's die opgedeeld zijn in 15 video genres. De dataset bevat ook talen die niet het Latijnse alfabet gebruiken wat inhoudt dat er wat data zal wegvallen. Maar de dataset is gigantisch dus zou dat niet uit moeten maken

6.2 Modelleren

Vervolgens ga ik beginnen met modelleren. Ik ga beginnen met het modelleren van een baseline model zodat ik die kan gebruiken als maatstaf voor mijn uiteindelijke model. Deze baseline ga ik trainen zoals in de paper "Bag of Tricks for Efficient Text Classification-[10]. Deze paper laat zien hoe je een efficiënte baseline model kunt trainen voor tekst classificatie. In de paper wordt besproken dat je een baseline model kunt maken door een bag of words model te maken en daar lineaire regressie op toe te passen.

Voor het daadwerkelijke model ben ik van plan de stappen op te volgen van Chollet in het boek Deep Learning with Python [3]. Dit is het boek dat we in de les hebben gebruikt op aanrading van de M&T docenten. In dit boek gaat Chollet voornamelijk aan de slag met N-grams. Dit is een zeer efficiënte techniek die ik ook heb gebruikt in mijn M&T opdracht waarmee ik erg goede resultaten heb behaald.

6.3 Evalueren

Als mijn verbeterd model een hogere score behaalt op de validatie set dan het baseline model is mijn doel bereikt. Ten slotte, als de validatie score inderdaad beter is dan het baseline model, ga ik beide modellen testen op de testset om nogmaals te zien hoe goed het model daadwerkelijk verbeterd is.

7 Structuur van het Artikel

Nadat ik de modellen gebouwd heb kan ik beginnen met de technische documentatie. Ik heb ervoor gekozen om mijn research te documenteren in de vorm van een tutorial / how-to-guide, gericht op software engineers. Ik ben zelf van aard een software engineer en in mijn studie was Kunstmatige Intelligentie geen verplicht onderdeel van de studie. Ik heb ervoor gekozen om AI te volgen als specialisatie, maar alsnog voelde het alsof er een barrière was tussen mij en AI, waardoor ik de behoefte had om tutorials te volgen die extreem simpel waren. Omdat ik deze tutorials zo erg kon waarderen wil ik zelf ook bijdragen aan de community door een simpele tutorial de bouwen die door software engineers, die geen tot weinig kennis van AI hebben, gevolgd kan worden.

8 Planning

Dates	Tasks / Deliverables
8–14 Dec	EDA op dataset
15–21 Dec	Baseline model uitwerken
22-28 Dec	Verbeterd model uitwerken
29 Dec - 4 Jan	Vacation
5-11 Jan	Vacation
12-18 Jan	Alles van voor de vakantie opfrissen
18–24 Jan	Tutorial schrijven
24-30 Jan	Uitloop

Tabel 1. Planning voor mijn onderzoek

Referenties

[1] Sonam Adinolf and Selen Turkay. 2018. Toxic Behaviors in Esports Games: Player Perceptions and Coping Strategies. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts* (Melbourne, VIC, Australia) *(CHI PLAY '18 Extended Abstracts)*. Association for Computing Machinery, New York, NY, USA, 365–372. doi:10.1145/3270316.3271545

[2] Nicole A Beres, Julian Frommel, Elizabeth Reid, Regan L Mandryk, and Madison Klarkowski. 2021. Don't You Know That You're Toxic: Normalization of Toxicity in Online Gaming. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 438, 15 pages. doi:10.1145/3411764.3445157

[3] François Chollet. 2025. *Deep Learning with Python* (3 ed.). Manning Publications, Shelter Island, NY. https://deeplearningwithpython.io/

[4] Wendy M. Craig. 1998. The relationship among bullying, victimization, depression, anxiety, and aggression in elementary school children. *Personality and Individual Differences* 24, 1 (1998), 123–130. doi:10.1016/S0191-8869(97)00145-1

[5] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41, 6 (1990), 391–407. doi:10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

[6] Sarah K. Egan and David G. Perry. 1998. Does low self-regard invite victimization? *Developmental Psychology* 34, 2 (March 1998), 299–309. doi:10.1037//0012-1649.34.2.299

[7] Minne Fekkes, Frans I. M. Pijpers, and S. Pauline Verloove-Vanhorick. 2004. Bullying behavior and associations with psychosomatic complaints and depression in victims. *The Journal of Pediatrics* 144, 1 (2004), 17–22. doi:10.1016/j.jpeds.2003.09.025

[8] Chek Yang Foo and Elina M. I. Koivisto. 2004. Defining grief play in MMORPGs: player and developer perceptions. In *Proceedings of the 2004 ACM SIGCHI International Conference on Advances in Computer Entertainment Technology* (Singapore) *(ACE '04)*. Association for Computing Machinery, New York, NY, USA, 245–250. doi:10.1145/1067343.1067375

[9] Gilbert R. Gredler. 2003. Olweus, D. (1993). Bullying at school: What we know and what we can do. Malden, MA: Blackwell Publishing, 140 pp., $25.00. *Psychology in the Schools* 40, 6 (2003), 699–700. doi:10.1002/pits.10114

[10] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, Mirella Lapata, Phil Blunsom, and Alexander Koller (Eds.). Association for Computational Linguistics, Valencia, Spain, 427–431. https://aclanthology.org/E17-2068/

[11] Jaana Juvonen, Sandra Graham, and Mark A. Schuster. 2003. Bullying Among Young Adolescents: The Strong, the Weak, and the Troubled. *Pediatrics* 112, 6 (12 2003), 1231–1237. doi:10.1542/peds.112.6.1231

[12] Noam Lapidot-Lefler and Azy Barak. 2012. Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in Human Behavior* 28, 2 (2012), 434–443. doi:10.1016/j.chb.2011.10.014

[13] Marco Lui and Timothy Baldwin. 2012. langid.py: An Off-the-shelf Language Identification Tool. In *Proceedings of the ACL 2012 System Demonstrations*, Min Zhang (Ed.). Association for Computational Linguistics, Jeju Island, Korea, 25–30. https://aclanthology.org/P12-3005/

[14] Master of Applied AI. 2025. *Automatic Text Simplification.* Master of Applied AI, Block 2 Report. Amsterdam University of Applied Sciences. https://github.com/School-Semester-Summaries/MAAI-S1/blob/main/Group%20Project/Project%202%20-%20Natural%20Language%20Processing/Project-Scope2025-26.pdf

[15] Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval* 2 (01 2008), 1–135. doi:10.1561/1500000011

[16] Ken Rigby and Phillip T. Slee. 1991. Bullying among Australian School Children: Reported Behavior and Attitudes toward Victims. *The Journal of Social Psychology* 131, 5 (1991), 615–627. doi:10.1080/00224545.1991.9924646 PMID: 1798296.

[17] Cuihua Shen, Qiusi Sun, Taeyoung Kim, Grace Wolff, Rabindra Ratan, and Dmitri Williams. 2020. Viral vitriol: Predictors and contagion of online toxicity in World of Tanks. *Computers in Human Behavior* 108 (2020), 106343. doi:10.1016/j.chb.2020.106343

[18] Amit Sheth, Valerie L. Shalin, and Ugur Kursuncu. 2022. Defining and detecting toxicity on social media: context and knowledge are key. *Neurocomputing* 490 (June 2022), 312–318. doi:10.1016/j.neucom.2021.11.095

[19] John Suler. 2004. The Online Disinhibition Effect. *Cyberpsychology behavior : the impact of the Internet, multimedia and virtual reality on behavior and society* 7 (07 2004), 321–6. doi:10.1089/1094931041291295

[20] Selen Türkay, Jessica Formosa, Sonam Adinolf, Robert Cuthbert, and Roger Altizer. 2020. See No Evil, Hear No Evil, Speak No Evil: How Collegiate Players Define, Experience and Cope with Toxicity. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376191

[21] Jing Wang, Ronald J. Iannotti, Jeremy W. Luk, and Tonja R. Nansel. 2010. Co-occurrence of Victimization from Five Subtypes of Bullying: Physical, Verbal, Social Exclusion, Spreading Rumors, and Cyber. *Journal of Pediatric Psychology* 35, 10 (05 2010), 1103–1112. arXiv:https://academic.oup.com/jpepsy/article-pdf/35/10/1103/2582483/jsq048.pdf doi:10.1093/jpepsy/jsq048