**The Application of Machine Learning to Encourage Human Contribution**

**Introduction**

In this report, I explore how machine learning not only allows machines to fulfil responsibilities that previously required human intervention but how machine learning can increase human productivity and contribution. I will first discuss the current state of machine learning, then study a case considering how Wikipedia, the world's largest free encyclopedia, has utilized machine learning to encourage contribution. Finally, I will evaluate the effectiveness of the technology's application, the threats it poses, and propose recommendations for its improvement.

**Background**

A recent study by Deloitte showed that 76% of business executives questioned thought that "cognitive" technologies would transform their companies within the next three years (Schatsky et al., 2017). At the forefront of the "cognitive" technologies, leading change in today's business is machine learning. Machine learning is a subset of AI that is dedicated to equipping machines to learn. Mitchell & Hill (1997) state that “Machine learning is the study of computer algorithms that allow computer programs to automatically improve through experience.” Applications of machine learning can learn and adapt without explicit instructions, and as such can fulfil tasks that previously required human intervention, such as fraud detection and targeted marketing (Tzanis et al., 2006).

Machine learning makes use of statistics to find and use patterns within large amounts of data to help make decisions. Machine learning algorithms can make use of many different types of data such as images, words or numbers (Chen et al., 2019). Where the digital revolution allowed developers to automate repetitive tasks that did not require human intervention, the machine learning revolution goes a step further, enabling machines to make decisions that previously only humans could.

Deep learning is a machine learning technique that can find even subtle patterns in data sets and can use such patterns to make predictions. Deep learning makes use of neural networks, which are similar to neurons in the brain, which, through training "learn" the ability to perform tasks. Humans do not to manually create the neural network, instead, they are created during the training process (LeCun et al., 2015). Unlike code, neural networks are essentially unreadable to humans at scale, which can increase the difficulty in verifying the behavior of the network in all circumstances.

There are three main categories that machine learning implementations fall under, supervised learning, unsupervised learning and reinforcement learning. Each category has a different approach to training machines. The development/ life cycle for all types can be generalized as in this graphic.

Figure 1: A Machine learning life cycle graphic (Google, 2019)

The development/ life cycle for all types can be generalized as in this graphic.

In supervised learning, data is manually "labelled" by the developers. This labelling process shows the machine what a correct answer looks like. The machine must then find patterns and relationships within the labelled data that it can use to predict outcomes from unseen data.

In unsupervised learning, the developers do not add labels to the data. Instead, they let the machine label and decide the relationships between data all by itself. This approach can lead to the machine finding, "hidden structures" in the data that a developer may not label, and this can allow for the machine to make decisions based on more nuanced patterns than supervised learning (Potentiaco, 2020).

Reinforcement learning comes from a completely different direction, Reinforcement learning utilizes a trial-and-error approach. The machine can create effective solutions after incremental improvements through the trial-and-error process. Reinforcement learning needs the developer to provide a mean for the machine to control the system and a mean for the system to give feedback on the machines attempt. As such reinforcement has been successfully applied to video games, where the machine is allowed to control the character, and the machine gets negative feedback from deaths and positive feedback from scoring points (Shao et al., 2019).

**Case Study**

Wikipedia is a free online encyclopedia that has been created and edited by volunteers. Wikipedia is tasked with managing enormous amounts of data. The quantity of content that Wiki manages (articles) is enormous. The largest-ever printed encyclopedia, the Yongle Encyclopedia, was 240 million words long. The English version of Wikipedia alone is 3.7 billion words long (Wikipedia, 2020). Further, Wikipedia operates as a non-profit organization.

I have chosen to study Wikipedia as an interesting case for the application of machine learning technology. Wikipedia poses the challenge of scale. Not only does Wikipedia require huge amounts of storage capability, but massive amounts of human labor (Wikipedia, 2020). It is this huge requirement for labor that means any task automation can have great benefits.

Being heavily reliant on the work of its volunteers (many anonymous), ensuring volunteers remain engaged is of paramount importance to the projects continuing success. Worryingly Wikipedia found that in the 8 years leading up to 2015, the number of active contributors decreased by 40%. Wikipedia believes that this disengagement was fueled by frustrating bureaucracy within the platform and toxicity from its contributors (Simonite, 2015).

Wikipedia has utilized machine learning applications to help combat both of these challenges (Wulczyn et al, 2017).

Wikipedia is a project built on the idea of freely sharing information and knowledge. Wikipedia aims to share impartial information, and contributors should strive for verifiable accuracy especially when the topics are controversial. Inevitably, contributors often disagree. As such, Wikipedia needed to provide a forum (Talk pages) for discussion of article changes. Unfortunately, these talk pages are often susceptible to the sharing of toxic comments.



Figure 2: An example of Wikipedia 'talk' page (Wikipedia, 2020)

Every article has an associated discussion page, this can be accessed by switching from 'article' to 'talk' at the top.

The sheer number of comments across the platform meant human intervention couldn't filter all toxic posts. Wulczyn et al (2017) state "The challenge of creating effective policies to identify and appropriately respond to harassment is compounded by the difficulty of studying the phenomena at scale".

**Evaluating how Wikipedia has adapted machine learning to facilitate changes in working practice**

Wikipedia wanted insight so they could explore the nature of the toxic comments left on the platform, for example, what type of user left such comments. Wikipedia and Jigsaw (an Alphabet company, under the same umbrella as Google) worked together to create a solution that could analyze, at scales previously impossible, the toxicity of comments. The developers combined crowdsourcing and machine learning to analyze personal left on the forums(Wulczyn et al, 2017). The work undertaken specifically builds upon the sentiment analysis and spam detection among others.



Figure 3: The model for supervised machine learning (Nvidia, 2018)

Wikipedia and Jigsaw utilized supervised machine learning techniques to help them achieve their goals. Labelled data is required for supervised machine learning. Crowdsourcing was employed to manually label a random selection of comments from the message boards. Wikipedia ensured that multiple people reviewed all the comments, allowing decisions to be made based on an aggregate opinion.

**Evaluating the benefits and drawbacks from the implementation of machine learning**

With the model trained, Wikipedia was able to analyze their entire backlog of comments, and in doing so, provide greater insight. Contrary to their expectations, they found that "the majority of personal attacks on Wikipedia are not the result of a few malicious users, nor primarily the consequence of allowing anonymous contributions from unregistered users (Wulczyn et al, 2017).

Wikipedia benefited greatly from the insight they gained, this insight empowered evidence-based decision making. Wikipedia knew a simple ban on a limited number of members would not resolve the issue and that rather a wider set of measures would be required. (Smellie, 2017). With this understanding, Wikipedia chose to apply the algorithm to all new comments, sorting them and to flagging those that would require human involvement. Flagging only the messages that need attention, alleviates the burden of mass surveillance from Wikipedia's moderation teams.

Here we can see two different ways machine learning has benefited Wikipedia. Machine learning has provided insight that can guide policy and has been used to proactively flag toxic messages.

As with all machine learning applications, the results can only be as good as the dataset provided (Low, 2020). Any bias in the datasets will become learnt behavior for the algorithm. In applying this algorithm at a huge scale, any bias that crept into the crowdsourcing would then be magnified across the entire Wikipedia platform. A risk here is that, messages may get flagged as toxic when they are not, or that toxic messages may get missed.

**Analyzing how Wikipedia has achieved greater efficiency using machine learning**

Unfortunately, as of 2018 (Wikipedia, 2018) state that "Experience of harassment has not declined since 2017 and appears to remain steady. Wikipedia's internal investigation found that there had been minimal change in community engagement since 2017.



Figure 4: The yearly change in community engagement and collaboration on Wikipedia (Wikipedia, 2018)

As such I have some recommendations for steps Wikipedia could take to further utilize machine learning to help them tackle this issue.

**Recommendations for Wikipedia continued use of machine learning technology**

**Bias Recommendations**

Wikipedia's algorithm opted to train its model using supervised machine learning. It is worth considering how different methods introduce bias into the resulting algorithm. Unsupervised machine learning techniques replicate the bias that is embedded in the dataset into the final algorithm, while supervised learning introduces bias into the data set during the data preparation process (the manual labelling of comments).

In each circumstance, the machine learns the underlying bias as a valid relationship, which it then replicates. Wikipedia and Jigsaw utilized the Crowdflower service to hire workers to label their data. The people hired by Crowdflower flower will not have the same experience as the moderators who have manually moderated the forums.

* I would recommend that Wikipedia try and determine if there are differences between the Crowdflower labelling and the moderators work, and then account for this.
* I would recommend training an unsupervised model and then comparing the results, it may that there is less bias introduced via this method.

**Regression Recommendations**

As time passes, the nature and language employed in toxic messages will change. The algorithms decision making, however, will remain the same because it still is based on the original labelled dataset. Therefore, it is likely the accuracy of the algorithm could drop over time.

* I recommend adding new manually labelled data to the dataset periodically and retraining the algorithm routinely.
* I recommend specifically adding any flagged comments that were not picked up by the algorithm to the data set. This would help the algorithm prevent the same mistakes being made again.

**Diversity Recommendations**

Wikipedia could also apply this toxic detection application towards tackling Wikipedia's diversity issues. With less than 10% of editors on the platform being female, Wikipedia does not have enough women contributing to the site (Paling, 2015). The lack of female editors is fueling a cyclical loop. Fewer female editors potentially leads to more content that is hostile towards women being shared and in turn drives away potential new female contributors.

* I recommend Wikipedia train a new (or variant of the) toxicity service to recognize content (articles as well as forum posts) that is hostile towards women specifically, and helps its reduction. Wikipedia would then find it easier to attract more female contributors.

**Algorithm Application Recommendations**

I have recommendations for how Wikipedia can alter their model, by utilizing machine learning technology at different stages throughout the process, to achieve greater results.

First, here Is a presentation of how the current model works. In the following diagrams I have highlighted the stage when action to prevent the toxic messages is taken.



* Before user posts: Wikipedia could notify users of toxicity in messages before the user posts them. Similar to how a grammar check may underline grammar mistakes, this check could highlight potentially harmful messages and explain. This provides an earlier opportunity for user intervention. This would change the model as depicted here: 
* After a user has posted: Wikipedia could display an icon indicating to other users posts that a post contains potentially toxic messages. This flag could be applied before a human has intervened.
* Block a user's posts: Alternatively, Wikipedia could block messages from being posted that fail the algorithms toxicity detection. This could however, annoy users if the detection was incorrect. This mild inconvenience may be considered acceptable compared with the benefit of blocking toxic messages before they are posted. 

**Summary:**

In summary, I have been able to analyze the current techniques being applied by Wikipedia and have provided recommendations in areas of Bias, Regression, Diversity and Application

.

**References:**

Chen et al. (2019) This Looks Like That: Deep Learning for Interpretable Image Recognition. Available from: <https://papers.nips.cc/paper/2019/file/adf7ee2dcf142b0e11888e72b43fcb75-Paper.pdf> [Accessed 09 Dec 2020].

Google. (2019) Machine Learning Life/Development Cycle Graphic. Available from: <https://cloud.google.com/blog/products/ai-machine-learning/building-ml-models-for-everyone-understanding-fairness-in-machine-learning> [Accessed 09 Dec 2020].

LeCun et al. (2019) Deep learning. Available from: <https://www.cs.toronto.edu/~hinton/absps/NatureDeepReview.pdf> [Accessed 09 Dec 2020].

Low, J. (2020) The Importance of Good Quality Training Data. Available from: <https://blog.supahands.com/2020/01/31/the-importance-of-good-quality-training-data/> [Accessed 09 Dec 2020].

Mitchell & Hill. (1997) Machine Learning. Available from: <http://www.cs.cmu.edu/afs/cs.cmu.edu/user/mitchell/ftp/mlbook.html> [Accessed 09 Dec 2020].

Nvidia. (2018) Machine Learning Model Graphic. Available from: <https://blogs.nvidia.com/blog/2018/08/02/supervised-unsupervised-learning/> [Accessed 09 Dec 2020].

Schatsky, D. et al. (2017) The 2017 Deloitte State of Cognitive Survey: <https://www2.deloitte.com/content/dam/Deloitte/us/Documents/deloitte-analytics/us-da-2017-deloitte-state-of-cognitive-survey.pdf> [Accessed 09 Dec 2020].

Shao et al. (2019) A Survey of Deep Reinforcement Learning in Video Games. Available from: <https://arxiv.org/pdf/1912.10944.pdf> [Accessed 09 Dec 2020].

Simonite, T. (2015) Artificial Intelligence Aims to Make Wikipedia Friendlier and Better. Available from: <https://www.technologyreview.com/2015/12/01/164848/artificial-intelligence-aims-to-make-wikipedia-friendlier-and-better/> [Accessed 09 Dec 2020].

Smellie, S. (2017) Inside Wikipedia’s Attempt to Use Artificial Intelligence to Combat Harassment. Available from: <https://www.vice.com/en/article/aeyvxz/wikipedia-jigsaw-google-artificial-intelligence> [Accessed 09 Dec 2020].

Paling, E. (2015) Wikipedia’s Hostility to Women. Available from: https://www.theatlantic.com/technology/archive/2015/10/how-wikipedia-is-hostile-to-women/411619/[Accessed 09 Dec 2020].

Tzanis et al. (2006) Modern Applications of Machine Learning. Available from: <https://www.researchgate.net/publication/228340464_Modern_Applications_of_Machine_Learning> [Accessed 09 Dec 2020].

Potentiaco. (2020) WHAT IS MACHINE LEARNING: DEFINITION, TYPES, APPLICATIONS AND EXAMPLES. Available from: <https://www.potentiaco.com/what-is-machine-learning-definition-types-applications-and-examples/> [Accessed 09 Dec 2020].

Wikipedia. (2020) Size Comparisons. Available from: <https://en.wikipedia.org/wiki/Wikipedia:Size_comparisons> [Accessed 09 Dec 2020].

Wikipedia. (2020) Community Insights/2018 Report. Available from: [https://meta.wikimedia.org/wiki/Community\_Insights/2018\_Report#Experience\_of\_harassment\_has\_not\_declined\_since\_2017\_and\_appears\_to\_remain\_steady](https://meta.wikimedia.org/wiki/Community_Insights/2018_Report" \l "Experience_of_harassment_has_not_declined_since_2017_and_appears_to_remain_steady) [Accessed 09 Dec 2020].

Wikipedia. (2020) Talk:Neural network. Available from: <https://en.wikipedia.org/wiki/Talk:Neural_network> [Accessed 09 Dec 2020].

Wikipedia. (2020) Community Insights 2018 Graphic. Available from: [https://meta.wikimedia.org/wiki/Community\_Insights/2018\_Report#Experience\_of\_harassment\_has\_not\_declined\_since\_2017\_and\_appears\_to\_remain\_steady](https://meta.wikimedia.org/wiki/Community_Insights/2018_Report" \l "Experience_of_harassment_has_not_declined_since_2017_and_appears_to_remain_steady) [Accessed 09 Dec 2020].

Wulczyn, E et al. (2017) Ex Machina: Personal Attacks Seen at Scale. Available from: <https://arxiv.org/pdf/1610.08914.pdf> [Accessed 09 Dec 2020].

**Bibliography:**

# Bernard Marr & Co. (N.D.) The Amazing Ways How Wikipedia Uses Artificial Intelligence. Available from: https://bernardmarr.com/default.asp?contentID=1549 [Accessed 09 Dec 2020].

Towards Data Sceince. (2019) Understanding and Reducing Bias in Machine Learning. Available from: https://towardsdatascience.com/understanding-and-reducing-bias-in-machine-learning-6565e23900ac [Accessed 09 Dec 2020].