Machine Learning Techniques and Applications in Social Media Analysis

Author: Marcelo Santos Pereira

INDEX

Introduction

This article is the summit of a research about a topic chosen by me accordingly with Machine Learning lessons' I've been attending for the 2019/2020 Academic Year second semester.

Starting with the main topic, Machine Learning (ML), it is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention (SAS, 2020).

As every technology wasn't always like we know them today, ML really started to rise proportionally to the rise of computer science since with the continuous upgrades to computers we could start to demand more and more of them which led ML to start being used everywhere nowadays.

As I wasn't aware and most us aren't, ML is used in so many different fields and one of them is social media. In the present date all of us have profiles in some sort of social media and many of us in multiple ones but we don't think about what's behind it and what information can a computer get just by analysing our profiles, our publications or tweets, our "likes", the people/pages we follow and our followers. All of this is data, all of this can and will give us answers if we so decide to but most of us, once again, don't think about it and this can lead to another topic that has started to get some attention over the last few years, cybersecurity and cybercrime. In this case also classification with machine learning is important as we can mark a potential profile or activity overall as concerning so that a team of people will study it saving thousands of hours. Usually this application of ML uses supervised learning. Not all of this is bad as this same data can give us a rough idea of what information is being consumed, for example, what pages get most views and why: is it getting searched? is it getting shared by the users? This helps with future marketing as well so that organizations can bet in different strategies to develop the interaction of people with their online profiles and websites that could get them potentially new clients and of course more sales if that's the main goal of the organization. Considering all of my previous research on the topic and all that I learnt from classes I decided to look for a real use of ML with a real scenario to get a deeper look and gather even more information.

Supervised and Unsupervised Learning

During my machine learning classes we spoke a lot about supervised learning which is when we teach the computer, with given a certain dataset, and we already know the answers. This dataset has the idea of pairing input with output and it is this precise matching that we need to teach the machine so it can potentially realize this process by itself.

So, like a professor would do, we correct them, and this is where the name comes from. We watch the machine and train it to the maximum efficiency. After we have a performance that we are satisfied with we start to make predictions using other datasets and that same methodology that the computer learnt. With the results of those predictions we can classified them. Also, in this model we label our variables. In these supervised learning problems, we categorize them as classification when the output are classes or groups (i.e. smoker and non-smoker) and regressions when the output is real fixed numbers (i.e. euros, height, weight) (Jason Brawnlee, 2016).

In contrast when unsupervised learning is used, there are no real correct answers (there's no fixed relation between input-output), we don't label variables and the algorithms are left by themselves to try and present us the data.

In this case, the output is treated in a different manner. Clustering and Association are used and the first happens when in the given data we already have groups and we try to group them according to some behaviour. The latter happens when we try to figure out fixed outcomes. For example, if person A bought X then it is likely to buy Y (Jason Brawnlee, 2016). Respectively, k-means and apriori algorithms are used with the already mentioned unsupervised learning problems.

Convolutional Neural Networks

In the process of researching the applications and usages of ML I also found that CNNs are present in many ML projects so I figure it would be a good idea to talk about it before looking to the applications and results of ML.

These CNNs are deep learning algorithms that are used for image analysis as they take images as inputs. Then it gives relevance to each image in the input and can learn to differ the said images. In this case is much easier for the CNN to classify the object than it is in the other classification methods, for example, supervised learning. Also, on a side note, it is important to say that before this methodology was developed the exact process had to be done by hand. This way we can train the machine to do it.

The name of the algorithm comes from the similarities that can be found between this same algorithm and what happens in our brain, the human brain. The main one is the having something happen and then the consequence, like a stimulant would perform.

Although I didn't mean to talk a lot about CNNs I feel like we should take a quick look to how does the machine interpret the input. Every image when read is associated with a matrix that has initially the size of the image. But it can also see the matrix of the pixels. So, with the classification process that I've talked about it can start to learn to recognize the images by itself. It could see certain patterns or something a little more detailed like textures.

When applying CNNs another part of it is pooling and this process is nothing but reducing and compressing the input (images). This way we have the same exact information but in a smaller scale. After we get the desired information and walking through every process that is needed, we get a dense layer that will give us the output.

In the end, there's no need to fit the data as pooling won't change anything it will only make it smaller in size.

Regarding the topic of the research visual analysis can be helpful to realize sentiment analysis in any social media or as I will later explain help to provide classification to any photo that is uploaded to a database by any user of the social media as it is the case with yelp.

General Machine Learning Techniques and Applications

I want to talk about a real scenario that happened in Spain but not before speaking a bit more about where is Machine Learning present in social media/applications.

As previously mentioned, ML is everywhere and to better understand this I went searching for real examples where companies/organizations have been keen to talk about it.

The first example is *yelp* as they created a photo classifier with the help of deep convolutional neural networks (CNNs) in the form of "AlexNet" (Wei-Hong, 2015) to recognize the classes that they labelled before. Their application is home for millions of photos uploaded by "yelpers", and this helps to categorize the photos upon upload. What I also found very interesting is that they have a report system that helps to improve their algorithm in case you find a photo which is not classified accordingly.

The second case is one of the most famous social media today, *Twitter*. *Twitter* used to show the tweets on our timeline in a very simple way. It would put together all tweets of the ones we follow and reverse them by chronological order which is neither very complex nor very accurate for what one would consider good content. So, they added a tweet score where the best tweets would show on top of our timelines but in order for us to not miss anything, they created a module called "In case you missed it". The features for the tweet classification were: (i) the tweet itself: its recency, presence of media cards (image or video), total interactions (e.g. number of Retweets or likes), (ii) the tweet's author: your past interactions with this author, the strength of your connection to them, the origin of your relationship (iii) you: Tweets you found engaging in the past, how often and how heavily you use Twitter (Koumchatzky, 2017). The former leader of Twitter-Cortex (one of their AI teams) also spoke about the introduction of deep learning, how they thought about it, their obstacles but mainly they made a better design of the platform and created a new training recipe for the data.

The next method I want to discuss is similar to twitter's previous example. Facebook, another leading social media also uses ML to rank the feeds, adds and search results but also to keep fake content away and to monitor ads as well. (Facebook Research, s.d.).

Getting to know this we can already see the impact that ML has on our profiles and on how companies deal with huge information. It is good practise to identify that both supervised and unsupervised learning is present in these examples.

Case Study

Since I spent so many hours studying and searching for supervised learning and cybersecurity because of 2 CU's this semester (Machine Learning and Ethics and Security of Information) I couldn't be more intrigued with what I found. In Spain, Bilbao, a methodology based on ML helped an elementary school to find the people responsible for cyberbullying.

The main idea was based on the fact that every trolling profile is followed by the real profile of the user behind the trolling one (Patxi Galán García, 2015). They also relied on the fact that we can detect authorship and with this in mind although the user would try to write in a different way, they could still find the similarities that would connect the fake profile to a real one. Having said this, they had only one question: It is possible to link a trolling account to the corresponding real profile of the user behind the fake account, analysing different features present in the profile, connections' data and tweets' characteristics, including text, using machine learning algorithms? (Patxi Galán García, 2015).

To answer to this question and before getting to a real-world scenario they had to develop their algorithm first. In addiction they had to restrain which profiles they would use and its conditions for being relevant to the test. With the help of twitter API, which is an interface that gives us all the possible operations we can perform from a said service, twitter in this case, they gathered the information. But it also had limitations, as they couldn't make more than 350 requests per hour. To surpass this problem, they used multiple accounts and only considered accounts with more than 100 tweets written by the owner so they could have a decent amount of text from each person. They would also consider the preferred platform of tweeting among other features as it would be crucial later to classify them.

For the classification of the tweets regarding the style of writing the team used WEKA.

Obviously, a few tests needed to be made in order to better understand the real performance of the methodology they developed. 1900 tweets from 19 twitter accounts (100 tweets per account) was the dataset used to test it. Vector Space Model (VSM) was the modelling tool chosen. VSM is an algebraic approach for Information Filtering (IF), Information Retrieval (IR), indexing and ranking (Patxi Galán García, 2015). The highest accuracy obtained was 68,47% using PolyKernel (maps the original dataset to a dimensional space and its short for Polynomial Kernel).

It was now time to test it in a real-world scenario and, as I already mentioned, the perfect case appeared when an elementary school had a growing problem related to cyberbullying. Its staff

asked if it was possible to find out who was the responsible person behind the troll profile, and the team got to work.

Initially the number of potential profiles was reduced to the class or the school as the facts being shared in the troll profile were super detailed. I also want to mention that that was a particular girl showed up more than any other person and even personal information was being shared on twitter. The full dataset had 17536 tweets belonging to 92 profiles which were followers or followed the said trolling profile and they applied their algorithm.

The final verdict from the system was that the profile wasn't being managed by one single person but multiple instead as it found similarities with 3 profiles that followed or were following the troll profile. The school with this new information appealed to the huge consequences if the responsible people didn't accuse themselves as they now had the possible 3 bullies. This was enough for the people to deliver themselves and the people that the team's methodology identified as guilty were the ones accusing themselves, proving the algorithm effectiveness. Although it wasn't reported to the authorities the needed measures were taken by the school.

Conclusion

This research really helped me to understand the importance of ML in our lives as well of how beneficial it could be both for better quality of a companies' product and for our own security as shown by the real cases mentioned for each side. Even if during the research I gave priority to topics regarding social media I couldn't help the curiosity that came with thinking about other close applications such as Google's voice assistant or Siri for example which use ML as well.

On the other side, in my personal opinion, a quick mention to the fact that today not every information we consume is based on what we like and what content we chose to follow but instead on what an algorithm thinks its best for us. Twitter's section "In case you missed it" is a step in the right direction as it still gives us the possibility to consume everything we chose to although in order to save us time it shows the tweets with better ranking first.

I will certainly look for more information about ML in other fields especially now that I know that even peoples' mental health can be preserved if we take into consideration the number of children that suffer from cyberbullying and how we can stop not only these situations but others regarding our and other's safety. I never thought the impact that ML could have.

Bibliography

SAS. (2020). From https://www.sas.com/en_us/insights/analytics/machine-learning.html

Jason Brawnlee, P. (2016, March 16). *Machine Learning Mastery*. From
        https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-
        algorithms/

Wei-Hong. (2015, October 19). *Yelp*. From https://engineeringblog.yelp.com/2015/10/how-
        we-use-deep-learning-to-classify-business-photos-at-yelp.html

Koumchatzky, N. (2017, May 9). *Twitter engineering*. From
        https://blog.twitter.com/engineering/en_us/topics/insights/2017/using-deep-learning-
        at-scale-in-twitters-timelines.html

Patxi Galán García, J. G. (2015). *Supervised machine learning for detection of troll profiles
        in twitter social network: application to a real case of cyberbullying.* Oxford
        University Press.

Facebook Research. (n.d.). From https://research.fb.com/category/machine-learning/