Sophisticated algorithms to prevent cybercrime

Using a novel AI model, a group of researchers from the Applied Mathematics and Computer Science Program of URosario successfully identified potential sources of cybercrime in social media.

By: Mauricio Veloza
Photos: Milagro Castro

On May 25, 2020, in Minneapolis (United States), African–American **George Floyd** was killed by a policeman. His death sparked a huge wave of protests in the United States and around the world that sought to expose police brutality and racial segregation. The slogan "***Black Lives Matter***" reinvigorated the huge movement against police racism.

Floyd's death was widely echoed around the world thanks to the spread of messages through social networks, especially Twitter, which was "flooded" with critical thoughts, accusations, and demands for political purposes.

"We took to the opportunity. Precisely when the *Black Lives Matter* movement was emerging, we found that most of the messages came from that social network, so the *hashtags* associated with that phenomenon became the main input for our research," said **Daniel Díaz López. "Twitter is a social network where nonconformity is exposed and multiple political positions are seen,"** he continues.

Díaz López is a professor in the undergraduate program and the **Master in Applied Mathematics and Computer Science** at the **School of Engineering, Science, and Technology** at Universidad del Rosario. He leads the study "**Development of cyber intelligence capabilities for crime prevention,"** (ongoing) which seeks to research advanced mechanisms for protecting cyberspace and preventing cybercrime.

In other words, its goal is to **analyze emerging technology to identify people who use it to commit crimes in cyberspace and ways to counteract these actions.**

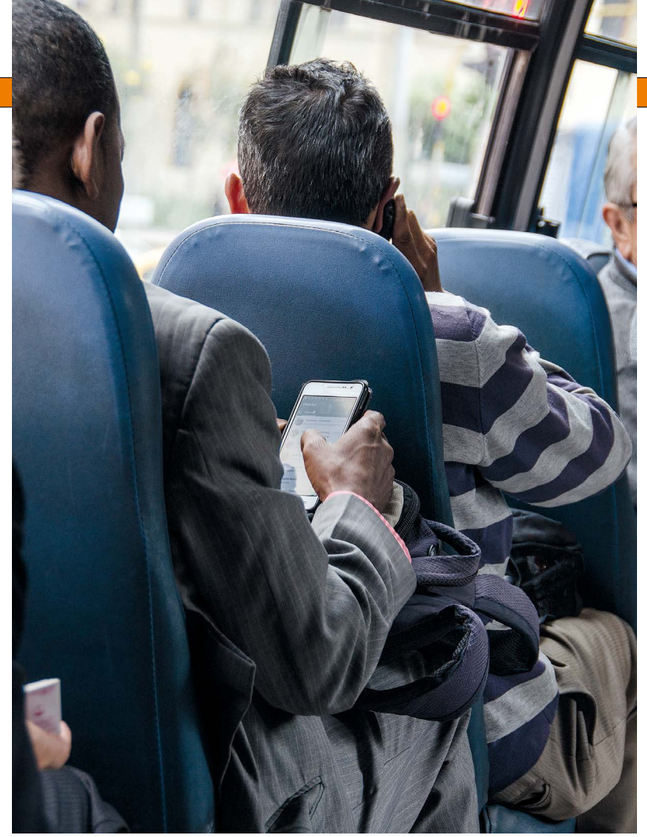
Social media is an important point in this research because it is an ideal place to express the messages of hatred and violence, which can lead to the commission of a crime. Moreover, because they are a "tribune" for misinformation, that in itself can be a punishable action.

The researchers analyzed 1,287 tweets in Floyd's case, but they also decided to look into other social mobilizations, such as those that took place in Colombia at the end of 2019. At that time, false information was published on Twitter, which generated panic and uncertainty in a large part of the population, as evidenced by a significant number of the 1,081 tweets analyzed in this case.

One of the cases of disinformation that was analyzed was the one in which the *hashtag* ***#DCblackout*** was used, which stemmed from an account with only three followers and became a trending topic shortly after. It spread false information about a widespread communications outage in Washington, D.C., causing major disruptions and many acts of violence in that city.

"There is a real and unfortunate situation: the Internet lacks sovereignty. Many times, people on social media feel they are more free or able to make threats or generate fake news. All that in a physical environment would be much more difficult to do because there are more social restrictions there. In the digital

Daniel Díaz López, a professor in Applied Mathematics and Computer Science at the School of Engineering, Science, and Technology of Universidad del Rosario, leads a research to build a solution based on artificial intelligence (AI), specifically for natural language processing, to support government agencies in preventing cybercrime.



scenario, people are "braver"; they are also protected by anonymity," asserts Professor Díaz López.

From his perspective, what is often seen on social media is that movements are generated for promoting hatred and violence. He says this is nothing but terrorism because it is understood as generating terror in society. Cyberterrorism, therefore, is the proliferation of terror in cyberspace.

"Social media has become a very interesting scenario to be analyzed because that is where movements such as white supremacists or those that seek to stir up massive sabotage against the security forces are born. That is where they grow," he says.

Therefore, he decided to conduct the research together with **Julián Ramírez** and **Alejandra Campo Archbold**, students in the undergraduate program in Applied Mathematics and Computer Science, and **Julián Aponte Díaz**, an officer in the Colombian Navy.

The goal of the professionals is to develop a solution based on artificial intelligence (AI), specifically for natural language processing, to assist government agencies in preventing cybercrime.

Broadly speaking, the research proposes using a similarity model based on natural language processing (NLP) to monitor suspicious activities on social networks. Through this model, a state security agency can search for posts that are similar, identify suspects, and thus anticipate the materialization or promotion of cybercrime.

This is how the similarity model works

NLP is an area of AI that seeks to build solutions capable of interpreting human language. For example, when a voice command is given to a cell phone, it has to recognize the spoken words. "This recognition function uses NLP, which is nothing more than giving a mathematical model the ability to understand what a person is saying," explains Professor Díaz López.

> Researchers will apply the model to specific topics such as photo-fines, panic attacks in finances, and any other topic that generates sensitivities and is expressed with certain emotions on social media, particularly on Twitter, where people's opinions are exposed.

Meanwhile, Campo Archbold comments that the research was based on the data science cycle: first the context and state of the problem were analyzed, then the data was acquired, then the modeling was created, and, finally, the deployment was implemented. "We retrieved the data using an application that recognizes tags and preprocessed so that we could use the similarity model afterward," he adds.

The first task they accomplished was the reorganization of the tweets: they discarded those with confusing or misspelled words, cleaned, and organized the data. Then, they did a vectorization, that is, they translated words into numbers. "This helped us to create a similarity model that allowed us to associate tweets. In the debugging process, we classified positive and negative tweets," says Ramírez.

In doing so, they used datasets with more than 500,000 words that indicated whether there was a positive or negative intention.

This way, they were training the model to identify which word was positive or negative. "There is always a margin of error, which is decreasing as you detect positive or negative intentionality," adds the professional.

Actually, two natural language processing models are used: one is the similarity model, which seeks to group, from the entire universe of tweets collected, those that have more similarities between them.

The second model is applied to those groups. "It is the feeling model, and it aims to detect the level of aggressiveness in each group. In the most aggressive groups of tweets detected, we try to identify their creators and replicators. That is the winning combination," says Díaz López.

Although it may seem strange **for data science to analyze human feelings, the professor assures that it is possible** because the data sets used are classified by people.

"This data set is used to train a mathematical model so that when we put a given tweet, the model encodes it and classifies it as positive or negative. The larger the data set, the more accurate the model can be because it will learn more. That is what we did for the second model," he says.

Manually performing this process would be very time-consuming; by reducing the time of analysis, a state agent's response time to detect where the hostile group or the focus of a probable cybercrime is can be shortened. In a nutshell, the model finds nodes that can impact the execution of potentially penalized activities and take quick action.

If there is a possibility of demonstrating that a plan to commit a crime was planned on social media, the perpetrators could be charged with cybercrime because they used cyberspace to promote a crime in a physical space.

A cyberdefense strategy

This opportunity for an early identification of possible cybercrimes provided by the model being developed by researchers in the Applied Mathematics and Computer Science Program led to the idea that it could be used by the Army.

"Clearly, it is necessary for the State security forces to control and monitor these types of dangerous situations to prevent crimes within the framework of the national strategy of cyberdefense," explains the professor.

However, as there is a risk that these messages on social media are misinterpreted, and instead of being the focal points of cybercrime, they are simply the impromptu manifestations of the rights to social protest, we need professionals trained to use the model.

"It is very important that data analysts can validate the data critically and objectively. No model of this type works independently; there must always be a human being who validates what the model says," notes Díaz López.

Another risk that could appear is related to crossing the line between privacy and autonomy for each individual. However, the researchers confirm that this line is established because there is an intelligence and counterintelligence law setting the boundaries of the government entities authorized to perform these kinds of activities.

Furthermore, the project included only open-source information, i.e., public data released on social networks and not private information.

NLP discovers the essence of words

Natural language processing (NLP) is the area of artificial intelligence that addresses human communication through computational machine learning models.

In short, it gives words a mathematical representation, where an NLP model could analyze the expressiveness of a sentence, interpret a person's desire from the use of certain words, or even establish similarities of intent between sentences.

Therefore, NLP offers a promising future for human language understanding, which can be useful in different fields such as customer service, advertising, voice translation, and suspect profiling.

Within the context of national security, it can be useful for detecting campaigns coming from hostile states and cybercrime organizations. Moreover, it may facilitate the resolution of cases related to disinformation campaigns against individuals or private companies.

"Hostile social manipulation" means the generation of violence and instability through social media.

The huge amount of information spread in this way makes it difficult to monitor and identify the source. For this reason, authorities are looking to data science as an ideal resource to collect, process, and analyze data, leading to the timely identification of such threats.

In other words, it is the same exercise that *marketing* companies have been doing for some time now to determine the impact of a new brand.

According to Campo Archbold and Ramírez, in the future, the research project plans to increase the features considered in the analysis of the tweets.

This will allow a deeper evaluation of the information obtained and the detection of advanced patterns of specialized threats.

Likewise, they will apply the model to specific topics such as photo-fines, panic attacks in finances, and any other topic that generates sensitivities and is expressed with certain emotions on social media, particularly on Twitter, where people's opinions are exposed.