

Early alert system for fake news

Fighting fake news in social media more efficiently and accurately:
Manuel Gomez Rodriguez of the **Max Planck Institute for Software Systems**
combines artificial intelligence techniques and human judgments to design
an early alert system for fake news.

Fact checking before the Internet era:
Pinocchio's nose gave away his lies.
However, this would not help prevent
fake news on social media platforms.



TEXT RALF GRÖTKER

Fake news is dangerous. In some cases, it is even life threatening. On December 4, 2016, a man with an assault rifle entered the pizza restaurant Comet Ping Pong in Washington, D. C. He had set out on a mission: to free the imprisoned and abused children that were supposedly held hostage in the restaurant. Just like millions of other Internet users, he had learned on the *Reddit* and *4chan* social media platforms that the basement of the pizza restaurant was the stronghold of a pedophile ring. At the center of the ring, so the tale went, was Hillary Clinton, presidential candidate at the time. Donald Trump's temporary National Security Advisor Michael T. Flynn and his son were among those involved in spreading this hoax.

The "Pizzagate" affair marks one of the high points of fake news to date. Various social networks have meanwhile begun to ask their users to report fake news. Some networks are also cooperating with journalist organizations that check facts. One example of this in Germany is [correctiv.org](#).

Manuel Gomez Rodriguez, Group Leader at the Max Planck Institute for Software Systems in Kaiserslautern, and his team are working on sophisticated methods for identifying fake news more accurately and efficiently. These methods are interlinked, much like the pieces of a jigsaw puzzle, and aim to analyze different aspects of the pieces of information that social media users receive in their news feeds, considering their respective context. "We are using a hybrid approach," explains Gomez Rodriguez. "We combine artificial intelligence techniques and human judgments to design an early alert for fake news."

Pope Francis shocks world, endorses Donald Trump for president



The claim that Pope Francis approved of Donald Trump being elected as President was shared by millions. However, it was a complete invention. This could have been revealed very simply: the website *WTOE 5 News* that published the news item refers to itself as a fantasy news website.

"Curb" has been presented by the researchers as a central result of their work. This algorithm is designed to prioritize which content should be most urgently checked and possibly marked as fake by a limited number of human fact checkers, signatories of the Poynter's International Fact Checking Code of Principles. The objective is to ensure that fake news are read by as few people as possible, before it is marked as fake.

A METHOD WITH A DYNAMIC THRESHOLD

A crucial aspect of the method is that it enables an elaborate analysis of the ways in which users handle content. This includes the frequency with which users share posts and the rate at which such content is then spread, as well as

the number of users that mark a post as fake. These are important criteria to estimate the speed at which a possible hoax will spread. Gomez Rodriguez: "While most social media platforms are currently merely choosing to fact check news items with more than a fixed predefined number of user complaints, our method uses a dynamic threshold that changes over time and reacts to the viral nature of a news item as well as the likelihood of it being a hoax."

More specifically, the algorithm developed by Gomez Rodriguez and his team focuses on the relation between complaints on the one hand and *shares* without complaints on the other hand. The more often a news item is proportionally shared without a complaint being made, the more likely it is that it is *not* a hoax. However: the faster a news

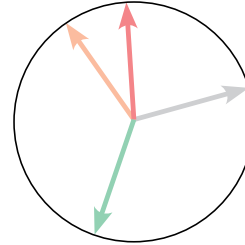
Comments

I like red candy!
I like green candy!
Candy is good and red is the best!
Candy is bad for you!

Voters

1	2	3	4	5
👍		👎		👍
👎	👎	👍	👍	👎
👍			👍	👍
👎	👍	👍	👎	👍

Opinions



Text analysis of user ratings: Max Planck researchers from Kaiserslautern analyze the degree of differentiation or polarization of statements made online, based on agreement (thumbs up) and disagreement (thumbs down) with statements that are part of a sequence (center). They use this information to determine vectors that are used to locate the statement within the opinion space (right). The illustration shows that the statements "I like red candy" (red vector) and "I like green candy" (green vector) represent opposing views. The researchers use the candy example to illustrate their approach.

item spreads, the greater the potential damage in cases where it is a fake news item after all. This problem is addressed by Curb by simultaneously monitoring and constantly updating information about distribution speed and the likelihood of a news item being fake. The algorithm's job is to optimize the balance between these two criteria.

For example: assume a news item is shared ten times an hour, and the likelihood of it being fake is fifty percent based on user ratings. It can then be mathematically concluded that, *on average*, five users an hour are exposed to a hoax. However, this calculation is now adjusted whenever a user shares the news item in question, and either *flags* it as a fake or does not object to it,

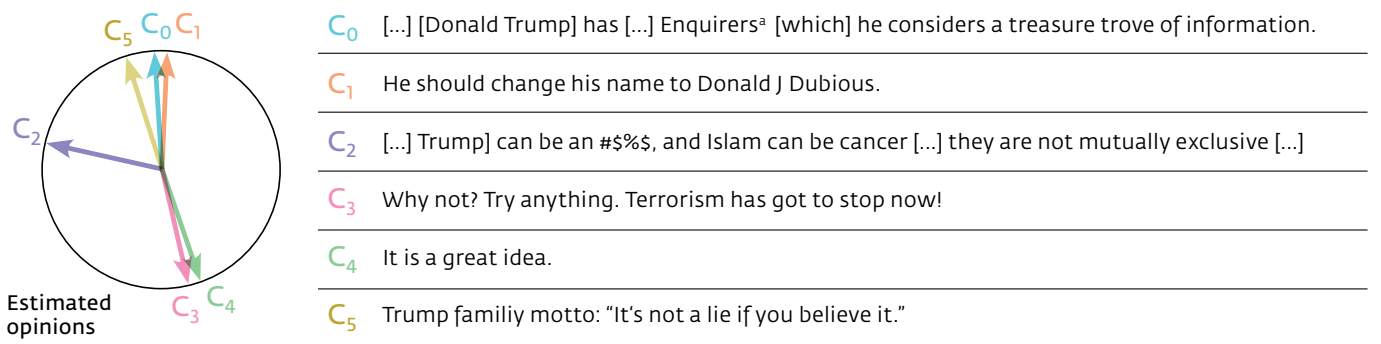
assuming it to represent a solid piece of information. The algorithm therefore dynamically creates an optimized balance between the endeavor to expose as few people as possible to undeclared fake news on the one hand and efficiency in using human fact checkers on the other.

TESTS USING DATA FROM TWITTER AND WEIBO

The ultimate test for Curb was an experiment conducted with real data that researchers from the Korean KAIST university had previously gathered from two popular social media sites, Twitter and Weibo, by means of web crawling, which they made publicly available.

The larger of the two datasets is the one from Weibo, with more than 4,600 individual news items, which 2.8 million users had posted or shared 3.7 million times. "We had information about the network structures in the dataset; that is the number of followers of each user, and we also knew which news items had been declared fake by the fact checking organization Snopes," explains Gomez Rodriguez.

What the researchers did not know was how and when the users from the dataset had flagged the posts. The researchers had to use a trick to solve this problem. They relied on other surveys about the number of times that users had flagged news that was actually fake. This enabled them to make well-



^a National Enquirers is a well known entertainment magazine in US.

Orientation guideline within the spectrum of political opinion: the way in which users assess the comments made in an online debate about Donald Trump provides an insight into the political view on which the comments are based, which can be recorded as vectors in the opinion space. Statements C₀, C₁ and C₅ were clearly made by individuals whose political view is entirely different from that of the individuals who made statements C₃ and C₄.

Vote Hillary from home! Save time & avoid the line!



The advertisement claiming that supporters of Hillary Clinton could cast their vote by sending a text message was intended to mislead voters. The advertisement bearing the Clinton campaign logo was distributed via Twitter. However, those who followed the appeal lost their vote, as it was not actually possible to vote by text.

founded assumptions about how effectively users recognize items of fake news, and how often they also mark them as such on average. “We simply made our algorithm try out a wide spectrum of plausible flagging behavior,” explains Gomez Rodriguez.

In the experiment with real data from Twitter or Weibo, the researchers from Kaiserslautern tested how effective their algorithm was in submitting suspicious posts for fact checking compared to other methods. The methods that Curb competed against included the pseudo-method Oracle, which in the testing scenario had access to the information as to whether a news item was a hoax or not, and which accordingly submitted the post for fact checking.

Other comparison methods employed simple general rules: an algorithm which – like the method created by the team from Kaiserslautern – determines the level of urgency of fact checking based on the simple ratio between the number of flags and the number of shares; another algorithm that submits a post for fact checking as soon as a certain number of flags has been reached; and finally, an algorithm that only considers the distribution rate of a post in order to prioritize a news item for fact checking.

FURTHER APPLICATIONS FOR THE CURB ALGORITHMS

The result of the comparative tests: Curb was almost as good as Oracle at preventing the spreading of fake news that had not been identified as such. The three methods based on general rules were less effective.

Despite the successful test, Gomez Rodriguez is not yet able to predict how Curb will be used in practice in the future: “It remains to be seen if, at the end of the day, Curb is worth considering as a solution, or if only certain components of our method will prove to be of interest to commercial providers,” says the researcher. “However, one of the developers of Curb has recently done an internship at Facebook’s fake news team.”

Algorithms similar to Curb can also be used in different fields. “Language learning software, for example, could be optimized by methods similar to Curb, by improving predictions of which content should be presented to the learners repeatedly in order to allow them to memorize this content,” says Gomez Rodriguez. Another area of application is the field of viral marketing. The basic structure of Curb was originally developed by the researchers for this type of application: to find out how posts can be spread most effectively in social media.

Nevertheless, there is one problem that Curb fails to solve: what happens

if users choose to sabotage the system, by flagging solid news items as fakes, or by deliberately spreading fake news? Such extreme behavior would make it hard for Curb to correctly assess how urgently a post should be submitted for fact checking. Gomez Rodriguez and his colleagues have developed “Detective” to address this very issue.

The Detective algorithm is also intended to reduce the distribution of false information. Gomez Rodriguez’ team presented the method at the Web Conference in Lyon this spring. While Curb considers all users to be equally reliable, Detective aims to identify users who object to fake news particularly reliably and those who deliberately mark solid news items as fake in order to undermine the system.

For this purpose, the Detective algorithm considers the results of fact checking and uses these to estimate the extent to which users are to be considered reliable in recognizing and flagging fake news. “We monitor a user over a certain period of time,” explains Gomez Rodriguez. “While doing so, we



Fighting fake news with artificial intelligence: the work performed by Manuel Gomez Rodriguez and his team includes the development of methods that allow for efficient prevention of the distribution of fake news that is not recognizable as such.

repeatedly submit posts they create or share for fact checking.”

However, the Detective algorithm is also subject to a conflict of interests. In order to be able to assess the reliability of the greatest possible number of users, fact checkers should, on the one hand, verify posts that have been shared by as many different people as possible. This includes posts that according to user flagging are unlikely to be hoaxes. They thus gather information about which users assess information reliably. On the other hand, here too, the limited number of human fact checkers should also be used only for posts that are likely to be hoaxes. The most efficient way to achieve this would be to simply trust the judgment of those users that were found to be reliable in the past. For further users to achieve this status, however, the machine learning techniques used by Detective must be exposed to the behavior of as many individuals as possible. Among the achievements of the Detective algorithm is its ability to find an ideal compromise between these two requirements by means of machine learning.

Just like Curb, Detective also passed the test with empirical data sets with flying colors. The results achieved by the method in the experiment were very nearly as good as those of a pseudo-algorithm that was familiar with the users’ flagging behavior. In practical application, Detective combined with Curb should be useful for administrators wishing to use the algorithm to allow for most efficient human resource planning in fact checking.

Based on the Detective rating, administrators would also be able to give users access to information about the reliability of other individuals within their social network when it comes to flagging posts as fake. “In reality, however, this is limited by data protection regulations,” Gomez Rodriguez admits. Many users even find it unacceptable if their “friends” or “followers” are able

to see which posts they like. “Marking a post as a piece of fake news can be just as problematic, because this often involves disclosing an aspect of one’s own political orientation.” This is why the results need to be suitably anonymized by Detective. “Ten percent of the reliable individuals in your network have flagged this post as ‘fake’: you could display this type of information,” says Gomez Rodriguez.

POLARIZING NEWS ON SOCIAL MEDIA PLATFORMS?

However, indicating that certain individuals are particularly reliable may also cause the opposite of the desired effect: users that lean towards conspiracy theories might choose to follow individuals who deliberately flag solid news items as fakes and who spread fake news – because they believe this information to be the truth that is withheld by the mainstream media. However, Detective proved to be rather robust when it comes to dealing with the deliberate distribution of incorrect information – especially thanks to the fact that the algorithm takes the users’ reliability into account.

In addition to the endeavor to effectively reveal hoaxes, Gomez Rodriguez’ team also addresses the issue as to what extent news items – whether they are fake or not – contribute to a polarization of views on social media platforms. The researchers have developed another algorithm to answer this question. This algorithm analyzes ratings such as “Thumbs up!” or “Thumbs down!” for text-based posts such as comments in online debates.

Instead of opinions regarding individual questions, the researchers consider entire opinion sequences. Gomez Rodriguez uses the following statements to illustrate this approach: “I like red candy!”; “I like green candy!”; “Candy is awesome, and red candy is the best!” and “Candy is unhealthy.” The respective view on which an individual com-

Paid fake protesters were bussed in to the anti-Trump protests in Austin, Texas.

ment is based cannot be reliably determined using software that, for example, analyzes text for certain words and compares it to other statements. This is not the case for opinions that are expressed by users, by agreeing or disagreeing to comments in such a chain of statements. The researchers analyzed these views expressed by different users, and on this basis were also able to calculate the view that is reflected by an individual comment.

When analyzing views that are reflected both in an individual comment and in the ratings for a sequence of statements, Gomez Rodriguez and his colleagues focused on two characteristics.

On the one hand they considered the degree of complexity, or the number of axes based on which the opinion space can be depicted. For example, if all participants of a debate hold either the same view or exactly opposite views regarding an individual issue, the answers can be sorted along a *single* axis. This type of debate is literally one-dimensional.

At the same time, the researchers also determined how far apart individual opinions are from each other. To do so, the attitudes on which the comments that make up the sequence are based are represented as vectors that form an opinion space. The respective vector is determined by the algorithm, by analyzing the way in which other users assess a comment. The arrangement of vectors provides information about the diversity of opinions. As Gomes Rodriguez stresses: "We are able to locate text-based posts that differ greatly in their semantic content, which comprise completely different words, and which may even include irony, in relation to each other within the opinion space."

An analysis of a large dataset from online comments on the pages of Yahoo News, Yahoo Finance, Yahoo Sports and the Yahoo Newsroom app has found that 75 percent of online debates take place on two or more axes in the




These photographs of numerous buses were the only evidence used by questionable news pages to support the claim that paid protesters were taken to Austin to take part in a march against Donald Trump. The buses were actually on their way to an event at a congress center located several miles from the starting point of the protest march. No evidence was provided of participants being paid.

opinion space, and are thus not held in a polarized manner. "This is a clear sign that debates on these online pages have not fallen victim to demagogues," says Manuel Gomez Rodriguez.

The algorithm therefore enables debates held on online forums or on social media platforms to be assessed, and the results obtained to date contradict the impression that such debates held in the anonymity provided

by the Internet are generally undifferentiated and are largely polarized by demagogues. As is the case with Curb and Detective, the algorithm shows that a hybrid approach using artificial intelligence and human assessments can help to promote objectivity in such debates. ◀

 www.mpg.de/podcasts/digitale-gesellschaft (in German)

SUMMARY

- A hybrid approach comprising artificial intelligence and human assessments can, in various ways, help to promote the objectivity of debates held online.
- The Curb algorithm prioritizes the urgency with which a post should be subjected to a fact check in order to prevent a possible item of fake news from being spread without being marked as such. To do so, it repeatedly reanalyzes how quickly a post spreads and how many users have flagged it as fake.
- The Detective algorithm is also intended to prevent the distribution of fake news, while at the same time takes into account the reliability of the users that mark a post as fake.
- A further algorithm analyzes the degree of differentiation of debates held on the Internet. It has been found that 75 percent of debates are not held in a polarized manner, which is a sign that the majority of users do not follow demagogues.