Data Journalism - group 5:

Alec Schweiger	11574151
Ewout Colijn	10624848
Julian Jancik	12357375
Melissa Quirijnen	11219432
Oriol Subirada Jimenez	11816899

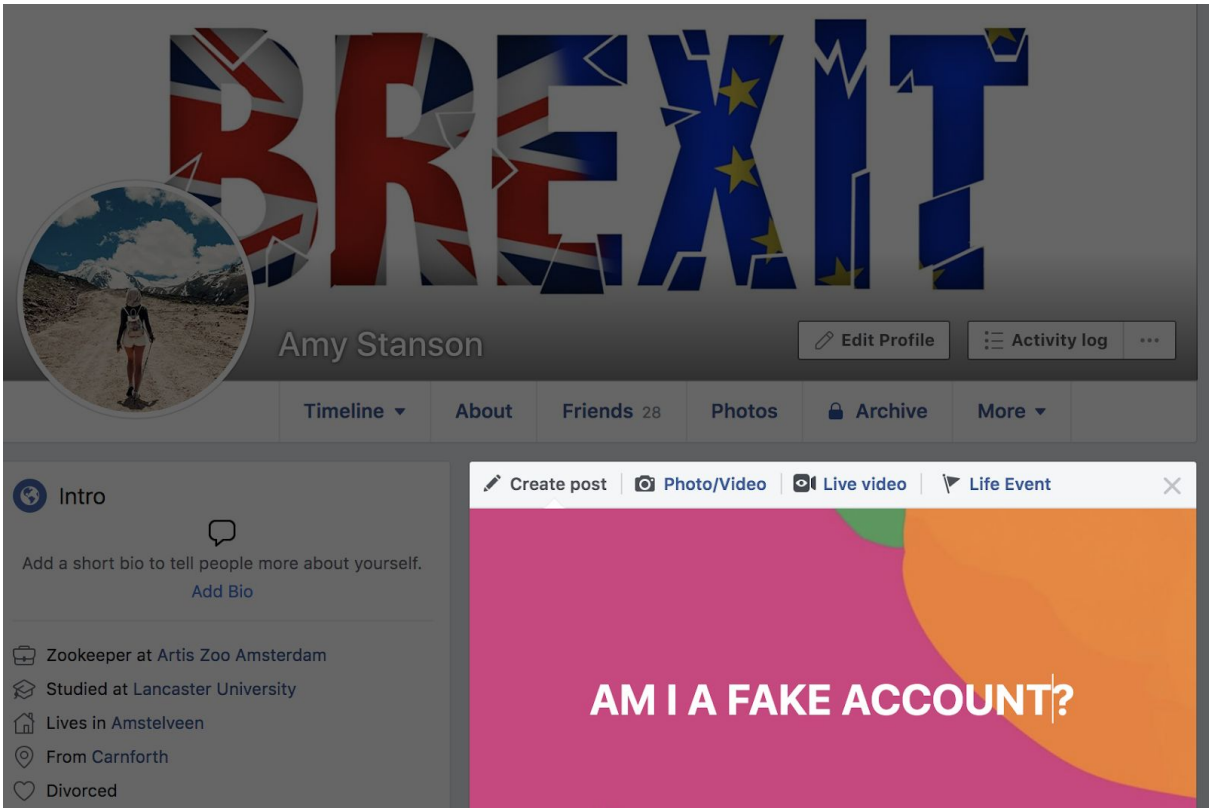
Final Report
Word count: 2641

Table of content

Becoming fake news for a greater good?

Amy Stanson is a 38-year-old expat living in Amstelveen. She is from the UK, where she studied at Lancaster University and where she formed her pro-Brexit and right-wing opinions. However the Brexit “mess” was too much for Amy at some point, so she decided that she needed a radical change in her life. The best solution she could come up with was to maintain her EU benefits by relocating herself to The Netherlands. She cut off most of her friends, figuratively like a lizard would cut off its tail when it flees for danger, except her five closest ones: Alec, Ewout, Julian, Melissa, and Oriol. After her relocation, she decided to focus on what she likes the most - reptiles. Occasionally she still shares and likes conservative content on her Facebook and from that point, she lived happily ever after.

This could be a wonderful life story with a happy ending, except for the fact that Amy does not exist. Her profile was manufactured by the Data Journalists of Group 5, yes, those best friends we talked about earlier, who decided to dive deep down into the dark waters of data journalism. With the help of Amy's data, they investigated the concepts of fake, real, algorithms, filter bubbles and more. So what did they find out?

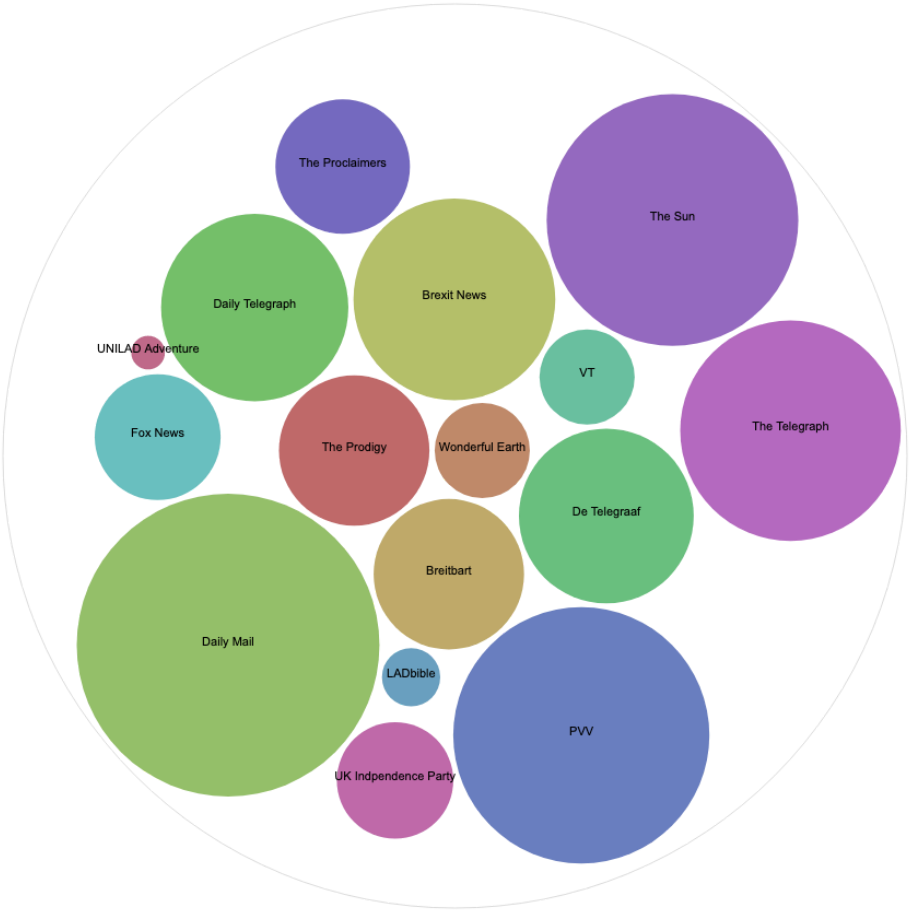


Outcome, result and the message from data

At the beginning of the project, Amy Stanson was not having any friends on Facebook. We observed, after collecting 1023 posts, with the help of the fbTREX tool, that her account only showed very one-sided content. The posts displayed were chosen after the principle you see what you follow. Even

though we were joining groups and we were liking and sharing content from other users, we did not get any friend requests. Nevertheless, Facebook showed us friend suggestions from people we all actually knew in real life. By the time Amy's news feed consisted of 0% advertisements, 0% posts from friends, 46,2% political and 53,8% non-political content. The amount of Brexit related posts were around 20%, which I will further elaborate on in the following paragraphs.

Shortly after that, with Amy's non-changing Facebook news feed, we decided to befriend Amy with our personal accounts. We saw an increase in 9GAG posts on her news feed since we were all following and liking posts from that page. The total percentage of it was 13.3% and the most represented account she saw. More radical Brexit news outlets went down to 14.6%, which could possibly be due to the very liberal characteristics of our personal profiles influencing her profile. Nothing changed in advertisement behavior. Nevertheless, the filter bubble did not burst. Since we all are not very active Facebook users we did not see a big change in her news feed. We decided to befriend more people.



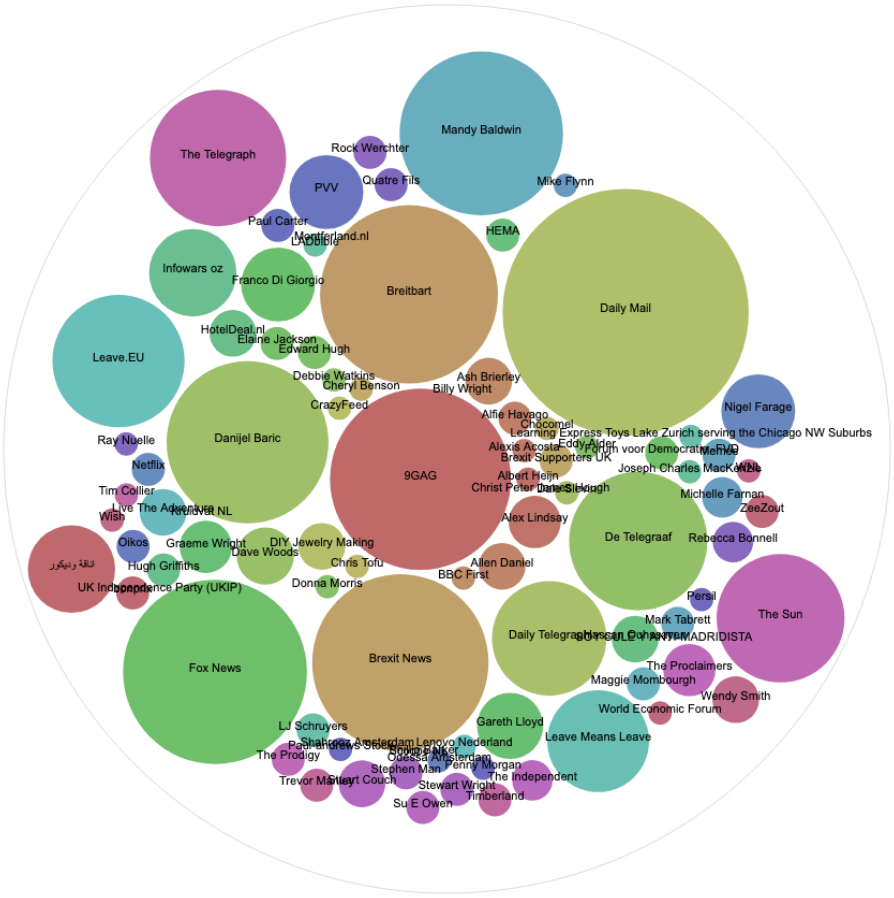
During the collection of the next dataset, we were having problems with the fbTREX tool. At several moments, when refreshing the Facebook news feed, we could not record more than 8 to 10 posts. Instead of "this post has been recorded" we got a notification that "fbTREX is not working for you, we want to fix this". After following the hyperlink we just ended up on the collection of Amy's data with no further instructions on how to solve this. However, we managed a way to get around this issue. After refreshing the news feed for more than 10 times we suddenly realized the posts were

recorded again. We noticed that at the time the posts could not be recorded, the news feed merely consisted of political non-friend related content. We just realized this because none of them included posts that could not be recorded from the fbTREX tool since they were all public and not from any friends. We experienced this issue two more times after that. We also realized that the tool could only store a maximum of 1000 posts in their collection, which can be limiting in bigger data journalism projects.

For the third data collection, Amy was in total having 26 friends on Facebook, that were ranging from Dutch accounts to Saudi Arabian accounts. The total number of posts we collected was 512. For the first time, ever Facebook showed us advertisements from for example Chocomel, Albert Heijn, Netflix or Timberland. These advertisements range very randomly and were not specified to Amy's own profile. Only the Dutch supermarket advertisements were, on the basis of her location in Amsterdam, specified. Friend suggestions now did not show us the people we actually knew in reality but this time it mainly showed us people from Arabian countries. This could be the case because nearly a quarter of our friends consisted of people coming from Arabia. In total 10% of the posts on Amy's feed now were from friends, 0,2% advertisements, 38,9% political and 61,6% non-political. One can see a big difference in political content since we now have more friends whose content was also shown. The posts related to Brexit were in total 22%, which is now getting important.

For the next dataset, we solely liked posts that were related to Brexit. We waited one day to make sure the algorithm has enough time to process the information. The fourth dataset consists of 839 posts.

What instantly got out attention was that most of the times 2 to 4 posts in the first 10 posts of Amy's news feed were Brexit related. After refreshing it for several times, we could be sure that it was not just a one-time occurrence. As well the 29 notifications Amy received on that day consisted of 17 Brexit related ones. Most of them were from Brexit discussion groups we joined already in the beginning, but which never really got attention from the Facebook algorithm. Also, posts on her feed included several posts from Brexit discussion group members, that made out 15,6% of the total amount of posts. Due to our higher number of friends we now received 2,5% advertisements. From the 22% of Brexit related posts in the earlier dataset, we now got to 30% after solely liking Brexit related posts just for one day. Posts from friends sunk down from 10% to 7,9% and 38,9% political posts turned into 41,3%. With this information in mind, we can now clearly think of the effect of a filter bubble. Amy's news feed got radicalized by Brexit related posts. Even if a rise from 22% to 30% does not seem enormous now we can definitely see this evolving in the future.



Real Persona vs. Fake Persona

To test the fidelity of our fake profile we have compared it to a personal profile, which has been active for over 10 years now. The data used for the analysis is a dataset of 1000 posts, which both have been scraped on the same day. It quickly became clear that the only viable variables for comparison were post-type, post source and the distinction between a business page or a personal profile as a source.

A first note when reviewing the data was the number of ads on the timelines of both profiles. For personal profile 18% of the recorded post were ads, however, for the fake profile, only 3,1% of the posts were ads. Interesting to notice is that before we started befriending people, our ad numbers were even lower on the fake profile, which might indicate that the interest for a profile by advertisers grows parallel to its social network.

The number of personal profiles and pages showing up on both timelines were similar. For both pages around 78% of the posts were from pages and the other 22% from personal profiles. The fake profile, however, scraped a lot of personal posts from Facebook groups that it joined earlier, while the personal profile had personal posts solely from its friends. These numbers suggest that Facebook tries to maintain a certain amount of personal information on a user's timeline. This could be a strategy to prevent people from alienating with their Facebook timelines.

The number of sources that we have recorded posts from varied heavily. The personal profile had 361 unique sources that made up the 1000 posts and the fake profile had only 91 sources that made up the 1000 posts. This shows that it is impossible to build an organic and layered profile in such a short time. The top 20 sources of the personal profile are varied and range from news outlets to personal profiles to hobby pages. The top 20 sources of the fake profile are solely news outlets or political pages. If anything, this shows that you can create your own filter bubble by feeding Facebook one-sided information about your views and what you like.

Newspaper Comparison

As an addition to our research, we investigated the posts of The Sun and The Daily Mail that showed up on our fake profile's newsfeed and the ones that did not. This seemed like another way to interrogate the choices that the algorithm is making for the profile. We have coded the content in five separate categories: location, violence, political, celebrity and viral. If the content matched one of the categories it was given a score of 1, if not it was given a score of 0. The posts used were recorded for 24 hours.

From the investigation of the Sun, the most interesting finding was that 50% of the posts of The Sun that showed up on our timeline contained violence, whereas the posts that did not show up on our timeline only had violence in 10% of them.

The investigation of The Daily Mail showed something completely different than the investigation of The Sun. From the posts that showed up on the timeline only 10% had violence in them whereas the posts that did not show up on the timeline had violence in 60% of them. The main portion of the Daily Mail posts on the timeline consisted of Viral content.

The comparison of the timeline vs non-timeline posts of the two newspapers above did not yield results that lead to concrete conclusions, however, they do show that there is a significant difference in the content that is chosen for a newsfeed and the content in general.

Ethical issues of data journalism

Data Journalism has become in quite a short period of time a remarkable force in journalism. When discussing data journalism, many focus on how to access data, how to transform said data into a story, and how to provide truthful information based on this data.

However, one important aspect of journalism is often put in second place to a certain degree when discussing data journalism; the ethics behind it.
Due to the fact that this kind of journalism often relies on vast amounts of numerical data gathered, interpreted and displayed, the focus is usually aimed at the numbers.

But one must not forget that, as in any kind of journalism, the ethics of it must be kept in mind at all times. This is necessary in order to provide truthful, accurate information and to protect the privacy of the people found behind the numbers, whilst maintaining the public interest.

A data journalist should at all times ask himself certain questions, such as what role does he or she played in the context of a certain article, what level of details are actually necessary to tell the story, and he or she should be able to provide corrections or updates to the data if required. Furthermore, it is important to check where the data comes from, how it has been gathered, and if it is already available to the public. If one keeps concepts such as these in mind, it should make for a more ethically responsible data journalism paper.

In our project, we were asked to create and flesh out a fake persona on Facebook, provide her with a certain level of personality, and analyze how the information presented to this account changed as time passed by and more context and information was fed into the profile. In experiments of this kind, we believed that ethics should play a major role when creating the profile. Since we were creating a fake profile, and trying to make it realistic, we had to keep in mind that to the public our Amy would appear (or we would try to make her appear) to be a real human being, sitting behind her smartphone or computer, just as any other user that would cross paths with her.

As soon as the profile was created, within our group certain ethical questions began popping up, such as if we should send friend requests to strangers, or if we should post or contact anybody on the platform with our newly made persona. These questions raised a debate within our group; on one hand, certain actions such as trying to befriend a large number of strangers could make for a more extensive, detailed report once we collected our data. In the other hand, however, there was this feeling that we may be tricking or misleading the people we were trying to befriend since they would not be aware that Amy was in fact 5 Media and Information students.

In the end, we decided to be as ethically responsible as we could with our fake online persona and stuck to what we believed were the most morally correct decisions. Due to this, we decided that we would not be befriending strangers and that we would keep our interactions with others to a minimum. We did this knowing that it may cause our data to be less effective to what we tried to achieve in our study, but in this way it felt okay and fair to keep up our fake profile.

Opportunities and challenges

During our project, we came across different opportunities and challenges. The main challenge was to break out of the filter bubble. We tried unfollowing pages but this did not do the trick. There was no significant change to Amy's timeline. In the beginning, Amy also did not have any Facebook friends at all. So the timeline only consisted of pages we decided to follow. This lead to a very monotone feed and also very monotone data. In an effort to change and broaden the data collection Amy got some friends. Unfortunately did also not have any effect on the posts in her timeline. In the end, we discovered that the opaque proprietary of algorithms had a big role to play in breaking out of the filter bubble.

This opaque proprietary of algorithms is problematic for data journalism in general since it is not clear how algorithms are designed and how they react to the data it collects (Diakopoulos 2015). We do not know the values algorithm assigned to certain data and how this is calculated towards the behavior of the algorithm as a whole. This lack of transparency explains why we had such a hard time breaking out of the filter bubble. We did not know which data to influence in order to change the algorithm which created a challenge for our research. Our lack of algorithmic literacy leads to our inability to break the filter bubble the way we wanted to.

Tools like FbTrex give the public the opportunity to become algorithmic literate themselves since it records the sources that the algorithm uses. Individuals with the right kind of expertise will be able to dissect the algorithm trough tools like FbTrex. Unfortunately because of privacy reasons FbTrex only collects public posts and not private posts. Since we do not know if the Facebook algorithm assigns different values to privates posts then it does to public ones, our research method may be flawed due to the opaque proprietary of the Facebook algorithm.

About the authors

The data journalism feature was co-authored by Alec, Ewout, and Julian. Alec played a crucial role in the extraction of the data and delivering the outcome/message from them. Additionally, he was responsible for visualizations of our data and results. Ewout's biggest contribution was to critically reflects on the data and finding interesting phenomenons/concepts and present them to the group for further elaboration. He was also responsible for the comparison between the Real and Fake Persona Julian was responsible for planning and logistics of the project, established the profile and wrote the introduction. The critical reflection of doing data journalism was co-authored by Oriol and Melissa, who contributed with outside the box points of views throughout the whole project and created some of the essential parts of the project's content. Every group member was responsible for curating our fake persona's Facebook profile and gathering the data.

References

Diakopoulos, Nicholas. 2015. 'Algorithmic Accountability: Journalistic Investigation of Computational Power Structures'. *Digital Journalism* 3 (3): 398–415. https://doi.org/10.1080/21670811.2014.976411.