ISOJ 2019: Day 2, Morning Session

The impact of artificial intelligence and machine learning on journalism

---

Chair and presenter: Nicholas Diakopoulos, assistant professor, **Northwestern University**

- **Clay Eltzroth,** product manager, **Bloomberg**
- **Lisa Gibbs,** director of news partnerships, **Associated Press**
- **Ling Jiang,** senior data scientist, **Washington Post**
- **Michael Morisy,** chief executive, **Muckrock**

Nicholas Diakopoulos: Hello, good morning. Welcome! Welcome to our panel on the impact of artificial intelligence and machine learning on journalism. It's great to see the diehards who are up early on Saturday morning. I'm Nick Diakopoulos, I'll be your moderator/host for the next hour and a bit.

We'll be discussing sort of everything A.I. and journalism, all kinds of interesting challenges and questions for us to get into on this panel. Whether it's machine learning techniques that help moderate online content which we'll hear about this morning or software that writes sports and finance articles automatically which we'll hear about from another panelist this morning.

I think there are some really fascinating and really kind of thorny questions for us to get into today things like what does A.I. mean for jobs in the news industry? How do journalists need to think about quality and accuracy of news content when it's being produced with A.I. and automation? And what are the interesting sort of ethical questions that come up as journalists integrate these technologies into their practices.

So, I'm going to be joined by a fantastic panel of experts today including Lisa Gibbs the director of news partnerships at the Associated Press among other things she works with industry leaders in emerging startups to identify smart applications of automation and A.I. at the AP.

I'm also joined by Clay Eltzroth as a product manager from Bloomberg one of the companies that's really gone all in on automation. Clay is helping to develop machine learning and speech to tech solutions for the Bloomberg newsroom. We'll also have Ling Jiang senior data scientists with the Washington Post where she works on data mining and knowledge discovery from large volumes of data.

And then finally we have Michael Morisy the chief executive of Muckrock where he oversees the development of several different technological tools used by journalists including things like Document Cloud, FOIA machine, oTranscribe and Quark pack.

So just a note on the format, I know we started a little bit late so I guess we'll have a shorter break in the middle. Each of us will speak for seven to ten minutes. We'll introduce ourselves, describe our work, our role, our projects as they relate to A.I. automation. Then we'll kind of get into tackling some of those big questions that I mentioned earlier, and of course we'll leave some time at the end for Q&A with that microphone cube.

All right! So, I'm going to kick things off and just tell you a little bit about myself and the kind of work that I do in this space. So, again I'm Nick Diakopoulos, I'm a professor of communication studies and computer science at Northwestern University. I director research lab they're called the computational Journalism Lab where we study design and build all kinds of computational algorithmic and automated production and dissemination tools for news information.

I've spent the better part of the last couple of years actually working on a book on these topics. It'll be out in a couple of months, I hope you check it out. It's called automating the news how algorithms are rewriting the media, but for today I kind of want to put a little bit of a of a kibosh on some of the A.I. hype that tends to circulate in the news media. I think when we talk about A.I. in journalism really what we should be talking about is human centered A.I. in journalism. And in particular I want to talk about three things that relate to human centered A.I.

One is that I want to talk about human values and technology and journalistic values in technology. Secondly, I want to talk about how humans and algorithms are increasingly hybridized or blended together in new news routines that incorporate A.I., and then I want to talk about how I think practitioners and scholars should really be adopting methods and approaches from human computer interaction when they're studying A.I.

OK, so first this idea of values and technology, so this shouldn't be surprising. All technologies including especially A.I. technologies embed and encode human values, these values reflect choices like what data was used to train the machine, how that data was defined and sampled, how algorithms are parameter arise and configured, how the defaults are chosen, what inputs the system pays attention to, what input the system doesn't pay attention to, what's quantified, what's not quantified, all of these are human decisions that play into how an A.I. system is designed.

Artificial intelligence systems are tools at the end of the day. They're built by people to serve human needs, and goals, and values, and that makes them fundamentally political tools as well. They express the values that designers and developers choose to build into them. I think this suggests a real opportunity for journalists and news organizations to become aware of and start exercising their ability to include their own professional ideological values, institutional values into these technologies that are then of course driving news production and dissemination.

If it's not going to be journalistic values going into these A.I. technologies it'll be values from non editorial stakeholders, values from platform companies and so on. I think it's time for us to start thinking about A.I. as a new medium, actually a medium in which journalists can express and exercise their ethical and normative values, their professional values through the code that they implement.

So I would call on journalists to kind of step up and really start being A.I. designers if they want to own the future of news media. So, to my second point hybridization you know... it's sort of a pattern in the evolution of media, every wave of new technology whether that's telephony, photography, the copy machine, digitization, all of these technologies somehow change the nature of roles and tasks and workflows in the newsroom.

A.I. is no different than any of these other technologies. It's also a technology that's going to change news work. It will often complement, but rarely entirely substitute for a journalist. Some economists have estimated that only about 15% of reporter's job and only 9% of an editor's job could be automated using current technology, current A.I. technology.

Humans still have an edge over non hollywood A.I. in two key areas that I think are essential to journalism. Complex communication is the first one and then expert thinking is the second one. So reporting, listening, responding, pushing back, negotiating with sources, and then having the creativity to compellingly put that altogether or knowing when the new angle of attack is needed, A.I. can do any of this stuff.

This is what we would call complex communication, but of course these are indispensable tasks to producing journalism. So, although A.I. can often augment human work to make it more efficient or more high quality more often than not I think we're finding A.I. technologies actually create new types of work. They create new tasks like configuring, parameterizing, knowledge management, data production, template writing. These are all new tasks that are coming online as a result of using these new technologies.

So, the future of A.I. in journalism has a lot of people around, but I think there's also a lot of research still to be done to study the new and changing news work that A.I. creates when blended into journalism practice. Now my final point is that you know whether we're talking about values in design or the evolution of hybrid workflows it's clear to me that the future of A.I. in journalism really has to be human centered.

I think to study this hybrid future scholars should be drawing on methods and approaches from the almost 40 year old field of human computer interaction, HCI, for instance value sensitive design approaches which have been around for dozens of years will enable the deliberate design and operation of A.I. systems within the value frameworks of journalism, and methods like task analysis which have been around for 30+ years will help decompose high level tasks into automated sub tasks that can be blended with human effort.

There are all kinds of new I think and exciting questions that can be tackled using these types of HCI approaches. So, just think about how should interfaces be designed to empower editors to edit at scale when an A.I. has just generated thousands of stories or how can human agency be maintained in these types of systems so that people can direct A.I. and uphold journalistic quality standards.

And how will end users ultimately come to interact with algorithmically imbued news media and A.I. agents. Again, I would argue that pursuing all of these types of questions will benefit from further collaboration with scholars and practitioners in the field of HCI. Something I'm happy to say we're already starting to make happen, the first HCI in journalism workshop will actually take place next month in Scotland.

So, I'm going to stop there, but I'm looking forward to hearing from our our panelists and then we'll jump into some of those exciting discussion questions.

Clay Eltzroth: All right! Well, I'm Clay Eltzroth. I'm a product manager at Bloomberg News and for some reason I always end up going after Lisa, and I could basically say "yeah, we do the same thing" but that would be just too simple to come up and do that.

But a little bit about my journey into the product manager role. So, I started off as an editor at Bloomberg. I learned very quickly that our clients care a lot about the money they invest in the news that we produce. I had a really bad run in with some of our clients and that all happened based on a headline that I set out based on a company called Sino Forest.

If you're familiar with Muddy Waters research and Carson Block that was the very first story that they had ever done and I got a lot of phone calls within I think milliseconds telling me a lot of bad things about myself that I didn't know and what that led to was the newsroom lawyer sitting next to me for the rest of time I was an editor which is also good because every single interview that I had, I had someone next to me to audit my questions and answers that I got from them in real time via email which wasn't fun either.

But that led to me starting a team within Bloomberg where we monitored for breaking news across all third party content including Twitter and at the time no one else was really doing it to the extent that we were, and I went to be a managing editor I helped roll out TikTok and then I had a great opportunity to come into a product manager role where I would help integrate A.I. machine learning and speech to text in the newsroom.

So I hopped at that opportunity, it's been less than a year. We've done a lot and usually you hear about what we do at Bloomberg, we hear the 20,700 journalists and analysts that we have, but what you don't necessarily here is we have 5,000 engineers that work with us and that kind of falls into a happy medium between more engineers than news organizations have, but significantly less engineers than big companies like Facebook or Microsoft or Google have and do all those engineers work on news? No, but they help gather data make processes easier for our clients, use the terminal and also for us.

So getting to automation, that's one of the reasons why I want to become a product manager. One of the byproducts of automation is a lot of content and thanks to that, it's great, you get a lot of coverage on things that you normally written quicker faster better, but you also get a lot of volume and also with social media you get even more.

So people are just inundated all the time with that and that's a big problem with the news. So, you've probably read the New York Times article on automation and what Bloomberg is doing. We use a tool called cyborg to process thousands of data points in seconds and send out headlines and generate stories, that's really cool. We used to have people that did that, they transition into roles where they're capturing the data, they're Q seeing the data and also looking through the press releases because automation is not perfect at least not yet for a lot of things.

And with the A.I. products that we've been developing I talked to a lot of journalists, I sit down, I explain it to them, and once I take them through the entire process they get why like "Wow, how does that happen? How does this model work so well?"

So, I told them that basically the secret is as we go through and we train the model during brief period time we'll go to the server and say "hey look, if you don't get this right we're gonna replace you the human" and I always tell a joke whenever I'm presenting things and a lot of people find it funny, but then there's always someone that's like "Oh gosh, computers are going to replace me" but that's not that's not really the case.

I think that like they've pointed out before Nic and Lisa reported rules are kind of transitioning. You can leverage data to make things quicker, faster, better, and I think that's very unique compared to when I first started as a reporter. Things are so much easier to do. Trying to sift through massive amounts of data would take weeks sometimes and now it take a couple of seconds thanks to A.I.

So, at Bloomberg what we've been looking at is how can we best use A.I. and that's to basically benefit the reporters and the people that are writing content and producing content. And it all boils down to one thing is what can you scale and for us if you're in the past, you're a reporter, you would hop on a conference call, you would take notes, you would try to rewind, but most of the time there's no place rewind during the conference call and wait on a replay that would show up two days later.

So the quotes that you would get and things like that didn't really work well, so we started transcribing these still involved one reporter one call. We tried reporters being on to calls at once had phones on each ear, that did not work out well. So, what we're looking at now is like how can we best utilize machine learning to bridge that gap.

There are thousands of companies that report earnings. We don't have thousands of reporters to listen to all those calls. So what we are all looking at is we can transcribe those calls and obviously the next step which I guess anybody would be doing is can we leverage machine learning to detect things that occur on those particular conference calls.

So yeah I mean it's something that I know that our clients do, there's FT article earlier this week about that. You know I think a lot of our competitors are probably doing the same thing, it's very interesting. One of the biggest things that we're doing at Bloomberg right now is we're going through and educating the reporters on what A.I. and machine learning is.

And I think that if we look at all the things A.I. machine learning can do. Getting the reporters and journalists on our side to help us do that is probably one of the biggest challenges for me right now because everyone's worried about the job, they think that the way they do things is very unique and they do it a certain way. So, I think the education around A.I. and machine learning is probably one of the biggest things I think for newsrooms going forward. And I think that's pretty much it. I'm really looking forward to the panel actually and getting a lot of questions from folks. Thank you!

Lisa Gibbs: All right! So, to set the stage for some of the big questions that Nic referred to, I'm going to walk through some of the projects that AP has undergone with regards to automation and A.I.

You know, when we first started automating company earnings stories back in 2014 we were one of the first news organizations, there was a handful of us that were experimenting with these technologies. Today there are news organizations all over the world, I mean dozens and dozens of examples of newsrooms that are using automation to make story production more efficient or algorithms to help us surface interesting patterns in data, and so this is really becoming more and more of a routine tool in newsrooms.

So, I just want to point out the use of the word augmented which you'll hear a lot and you heard anyone who was in the Microsoft session this morning also heard I mean we absolutely believe as Nic was saying that this technology is going to help journalists do their jobs better if executed properly of course. It's not going to replace the work of journalists.

So just again like when we think about how we are going to use these technologies, we look for a few key criteria. Is this going to remove routine tasks from the work of journalists so that they can do more creative and high impact work? Are these tools going to help us break news faster or find news in areas that we haven't searched before? Can it allow us to produce content that will enable us to reach new audiences or new markets?

And at AP In particular we have kind of a side benefit that not all newsrooms have which we have chosen as our strategy to identify startups with expertise and things like machine learning and linguistics to work with and that brings new skills and energy into our newsroom and helps us innovate faster. Probably as we'll discuss later I mean other newsrooms you know hire data scientists in their own and build things in-house, we have pretty much chosen not to do that.

So when you think about how to apply these technologies I mean break down what we do newsgathering, news production, and news distribution as a large global news agency I'm really going to focus on the first two aspects of this. And I'm going to talk first about the second part of the process of news production because that's where AP started and this and it's where we're best known.

So the fundamental problem we are great at identifying stories and writing them, but how do we do that at scale? How many stories can one journalist write and how many would we want them to write? So, especially for routine things like sports stories or earning stories are there ways to scale production of those that save time. And so in 2014 this was the project that really kind of put us on the map with this.

I was business editor at the time which is how I fell into this whole world. I had editors and reporters that for one month every earnings season so quarterly I mean pretty much all work stopped and everybody focused on producing cramming jamming out their earnings stories and what I often say is my reporters and editors felt like robots doing this work quarter after quarter and it wasn't incredibly exciting. So, we looked at how can we make this a more efficient process. We worked with a data vendor, we worked with a software vendor called Automated Insights and we wrote templates to automatically generate those stories.

And as you can see on the slide we're now automating approximately thirty seven hundred earning stories each quarter, most of the U.S. stock market. I'll mention that what we were able to do in that project was viewed as so successful in terms of freeing up our journalist time. I mean I as business editor I was able to reassign reporters to cover technology, personal finance, more breaking news. I can draw a direct line between automating earnings and the number of investigative stories my business department produced.

So, it had a real impact on the quality of journalism that AP was able to produce out of its business department. Just really quickly sports has been a real growth area for us. The demand for sports previews and recaps because of fantasy and sports betting has only grown and we're able to serve these needs more efficiently through text automation.

This year 2019 I should say by the end of the year we will be producing 40,000 automatically generated stories. Now that sounds like a lot, but remember AP global news agency, we produce about a thousand stories a day. So even though 40,000 sounds like a large number, it's still a very small number in the context of what we do.

Other news outlets are doing a lot with automating election results. Things like home prices you know, anything really that has very structured data can be automated. One of the more interesting projects that we're working on now and I should have a slide on this, but I do not. We are working to see if AP can't use its automation skills combined with local data. State and city level data to help local news outlets create more content that's relevant for their news audiences.

Many of the smaller news organizations don't have in-house automation expertise and so we're working on a project right now with the Newsday Media Group and Columbia University data science students testing around education data in the state of New York to see how can we work together to help local news outlets use automation to serve their audiences, and that's a project right now that we're really excited about.

Another project we're really excited about like I mean it's one thing to produce thousands of stories about you know basic home prices or earnings, but we're using natural language processing to create summaries of all AP stories. And what this summarize or does is it takes a story that was written for you know your basic digital platforms as we all would write them, we write in past tense, we use quotes, you know we do all of the things that journalists do and it creates a two to three line summary that's in present tense, no quotes and it's meant to be spoken.

For us this is important because those summaries go to broadcasters and radio customers as well as Amazon, Alexa and other voice devices. So, there's a business case for us to use this kind of summarization, but we are really excited about it because it's kind of our first experiment with how to use natural language processing and we think that it's the beginning of our foray into versioning you know, can we use NLP to create different storytelling approaches that may serve audiences in different ways. So, this is something that is a real big experiment for us and we're really interested in how it turns out.

You know, there are other ways of course that newsrooms are automating production. I mean video transcription which you heard about in the Microsoft session this morning is a big example. There is a machine learning assist to it in that when you're correcting video transcripts the machine learns from its mistakes and gets better and better. So that's another big time savings tool that we use and many others.

But I want to move because I'm worried about my time, 2 minutes. I do want to talk about news gathering which is just a really incredibly exciting area. So, event detection I mean that's a fancy word for scanning the sea of social media out there and detecting what of all these posts is likely to be a newsworthy event. So, we just finished a five week test, this tool is from a company called Sam and we were looking to see, can this tool help us break news faster? And it did.

Over the course of the five week test, we had 27 testers only 10 responded to a survey, so I've got to go crack the whip, but those 10 identified 50 instances where Sam told us about the news event before our editors knew about it. So, imagine the power of being able to find out about the Aurora workplace shooting or the fire in Brazil, you know even 10 minutes or more before you would have otherwise. And so, how these algorithms can help us identify news happening in our communities it's just a fabulous area of exploration.

I don't think I have a slide on it, but I do want to mention its cousin which is insight detection and there are companies like crazy Kraszna and their CEO is here in this room somewhere and others who go out of M.I.T. media lab that are really interested in exploring how to use social listening technology to help us detect insights particularly in underserved communities or underserved areas so that we're not just getting our news from the people who are the loudest on social media. And we've worked with both those companies to play around with how to our beat journalists who are covering immigration along the border or statehouse news in Illinois or whatever that might be, well how can they use algorithms to see what people are talking about in their communities around these topics.

So I think this whole area is really promising and a fascinating one right now. The last thing I'll talk about I swear Rosental. Authentication, now we happen to be building a tool called AP verify that helps us identify video that is factual and authentic and kind of how this works so the system will allow us to pace the U.R.L. and it all use machine learning to go through all of these processes that help speed the process of UGC verification.

So, it'll go out there and search on the Internet to find hey, has this video showed up before? Maybe it was two years ago, maybe it showed up as something else. That gives us a flag that this is not what it's meant to be. So I think the use of verification and that sort of thing are also really powerful uses. What else do I have here?

All right. I'll leave you with this final word "Bad Data, Bad Story" L.A. Times quake bot tweeted that there was a 6.8 earthquake in Isla Vista. It happened in 1925, and 1984 on Lifetime said: "Tell him not to be too hard on himself, we all make mistakes" thank you.

Ling Jiang: My name is Ling Jiang. So, I am a data scientist at The Washington Post so today I like to talk about how we leverage A.I. in the newsroom and the Post. So, we built lots of data driven products using machine learning techniques to assist the journalist. We hear lots of positive feedback about how machine learning makes things more efficient and effective, but we do also face lots of challenges.

So, the biggest concern is the reliability of the machine learning tools. The fact is that the machine makes mistakes. Well human does too, but the machine cannot correct itself without human intervention. If we use a robot moderator to review online comments can it always make the right decisions?

Obviously the answer is no because there is no perfect model that will be always 100% accurate in the real world. But what are the possible consequences if those machine learning tools are making mistakes in the productions. Actually it depends on the use case. For example, take the robot moderator as an example again, so mistakenly approving a common is not good because we want to maintain a civil conversation in our comment section, but mistakenly deleting a comment is much worse.

We don't want to push our commenters away by removing their comments for no good reason, but some other systems like for example headline or generation system will have a much higher tolerance for mistakes because the human will make the final decision before taking any actions. Last but not least just like many others in industries that start using A.I. journalists may wonder if A.I. is the threat to their jobs.

So next I will talk about 2 use cases at the Post and then and discuss how we deal with those concerns. At the Post we have about 2 million comments per month. So, commenters are really great critical to us because they are engaged in the royal users, but the challenge is that some of the comments are really hateful and if we left the comment section untended the trolls will crowd out thoughtful discussions.

But the question is a manual moderation is very expensive and actually infeasible at such massive scale so the result is that some news organization is shrinking their comments section or even outsourcing to other platforms, but to the Post reader engagement is too critical to curtail or outsource so we turn to the A.I. and the ModBot is how we automatically moderate the user generated content.

So ModBot works by giving every common score between 0 and 1. So the closer to 1 means ModBot is more confident in deleting the comments and closer to 0 means ModBot is more confident in approving the comments. So we can see we have 2 common example comments here. Comment A is really bad so ModBot gives a score close to 1 and for another good comment B here it gets a very low score.

So, the challenge of deploying a system like ModBot is we need to develop trust with the users. In the production we use threshold to automatic moderator comments with ModBot. So, for example here we use 0.8 and 0.1 for all thresholds so everything above 0.8 will be automatically deleted and everything below 0.1 will be automatically approved and every comment left in between will need human review.

So, in this case ModBot is pretty accurate in the prediction, but there is still a big chunk in the middle section that requires a manual process so we can adjust the threshold by changing to 0.7 and 0.2, but in this case we might introduce errors. So, we give our comments moderation team the ability to control thresholds of when to trust the ModBot versus when to require human moderation so they can change the threshold in order to decrease the amount of manual process as they gain trust with the system.

So as I mentioned before mistakenly deleting a comment can be much worse. So they can choose to be more conservative in the deletion and to avoid potential risks. So we can see that the Post of moderation team used to spend most of the time reviewing comments based on some online discussion policies so with ModBot taking care of the wrought work they can now have more time to focus on more complicated work such as how featuring high quality comments or interacting with the users.

So we can see that ModBot is actually assisting people instead of replacing them because most machine learning tools are trained based on historical data so they are actually mimic what humans did in the past and repeat the patterns so they can never replicate that creative part of the human jobs. And with ModBot our readers can have more space to express their thoughts in the healthy comments section and our journalists can have more time now to engage with the readers.

So, all these changes can be very beneficial for growing community and improving the reader engagement. Another example is the article popularity prediction tool. So, we have over a thousand pieces of news content every day, but not all of the article will be equally popular.

So what we need though the strategy for editors to prioritize the rising articles so they can enrich the content to increase the reading quality. Basically the system consider four different types of features for articles like metadata feature including article type section, contextual feature, sentiment readability of the text, temporal feature like the first view since publication and the social features like the in the 30 minutes tweet volume and so on.

So we built a regression model trying to predict the page view of the articles in the next 24 hours after publication because by its nature the lifespan of the narco is very short. So it is more interesting and valuable to predict an articles early popularity instead of the long term popularity. So if the system says an article will receive over 100k views in 24 hours we will push it out to a select channel so the newsroom then can allocate resources to improve the quality or enrich the content by adding images, videos or contextual links.

And actually in a system like this has a much higher tolerance for mistakes because the added risk will add their own judgment to the prediction before they decide whether you invest a resource to the articles. So with this popularity prediction tool we can prioritize the rising, our articles and optimize the resource allocation. In this case our editors can put more effort into increasing the content quality so we can serve better reading experience to the broader audience and also from the business point of view we can support advertising opportunities.

So we can see that the A.I. tools really increase the efficiency of our work because they can tend to the simple and repetitive words so free journalists to focus on the high value work and it also gives us the capability to deal with some tasks that are very massive scale which can be actually impossible with pure manual efforts.

And the journalists can also have better communication between their peers and with the audience and that increases the reader engagement. So one last thing I want to say is that technology is neutral. The key is how we use it and with appropriate strategic quality control and human intervention, we can make the best out of the A.I. tools. Thank you.

Michael Morisy: Hi, I'm Michael Morisy! I'm the founder of Muckrock foundation, we now run a variety of websites as we were talking about earlier. Our big ones are Muckrock which helps you to file track and share public records requests throughout the United States. We also run Document Cloud which thousands of newsrooms use to host, analyze and share their documents with their readers.

We also run oTranscribe which helps with transcription, FOIA machine which is sort of a self tracking public records tool. We also run a QuackBot which is a slack chat bot that helps with common newsroom tasks, and then we just launched our sixth service which helps manage your account across various different services, we got less creative with naming things so we just called it Accounts.

But overall we work with about 2,000 newsrooms and about 10,0000 users primarily journalists in those newsrooms and we reach about 20 million people across sort of through our newsroom users, and so we've been really working on some really exciting technology and sort of been trying to figure out sort of what's the next big thing when it comes to helping our users to understand their world and then present that information to our users.

And so the exciting technology I want to talk about most today is something new called PDF's, and so can I get a quick show of hands... has anybody's newsroom used large piles of PDF's or gone through large sets of documents? Okay, great! Keep those hands up for the next seven minutes because I want to follow up with you afterwards.

And also who has had challenges going through those large sets of documents? Okay great. I was hoping that nobody would raise their hands and I could just sit down and we could save that time for later, but pretty much all we do is go through government documents, many in the form of PDF, many in the form of emails, we've helped file through Muckrock, 60,000 public records requests so far and on document cloud we host a hundred million pages of primary source materials on behalf of our users.

So we have a lot to to kind of go through and so we first started getting interested in machine learning because of Muckrock, so when you file 60,000 request not only does that mean you have 60,000 items to kind of go through, but you actually have hundreds of thousands of emails with various government agencies that all need to be tracked and categorized. With Muckrock what we like to do is take as much of the drudgery of the public records requests process from journalists so they can focus on their reporting and that means we take on that drudgery.

And so that means going through hundreds of thousands of emails like that being like "hey I'm still working on this request it's going to take more time," we need to go through and manually classify all of that information, and there's one of eight possible cases that each classification can be some of them are very simple like this, some of them are a little more complex and on average they took about 8 minutes per each on average.

So this one might take a minute, another one might take 10 minutes as you kind of decide for the bureaucratese, but overall it was taken about 8 minutes per task. So if you total add up and I think we had three hundred and thirty six thousand different of these communications had to go through, they take about 8 minutes each, that's about 24 years of somebody working on this full time which wouldn't be too bad except at the time we only had 3 staffers and we had 0 staffers who were injured and working on this full time. So we decided to look for a better way.

So what we did was we built a micro task system, there's got to be a better name for this stuff, but we built a micro task system similar to like Mechanical Turk except only for internal use to streamline that process and make it easier to kind of go through tag this stuff. We had some of our staff members who would say you know what this is kind of soothing, I've got something on Netflix, I'm doing the categorization work, life is good.

And this actually cut down the time on the task just this alone cut down the task time 50% and also made it really easy to sort of onboard new people so if we had an intern who had some extra hours we have them working on this with just a few hours of training and it really kind of smooths out the process. So now we've gone from 24 years of cat task categorization time to just 12 years which is still more than we had people, but we're going in the right direction.

But if anybody has experience machine learning you can kind of tell that sort of in the back of our minds the goal of this sort of more precise tagging was so that we could automate this process going forward. So you can see OK, we've now got everything in a nice database, we've now got everything sorted categorized, we now make it so that we can have lots of people kind of digging into this information and make it look a little smoother.

So after a roughly one hundred thousand categorizations so that's a lot of back end categorization work, but eventually we got a model that was working about 80% of the time which was actually really exciting and when our lead developer is like all right, I've got an 80% accuracy I'm like great, now you've limited 80% of the work, everything will be good, I can just you know do this over breakfast once a day and we'll be all set.

But it turns out sort of if you are wrong when you follow up with the user one out of five times that user is not going to trust you very much. If our news articles one out of five of them was just completely wrong people probably won't come back to our news organizations or maybe they would these days, but we wanted to do sort of a better experience for our users and so we put a very high degree of sort of certainty with that model to kind of make sure that we are giving them really good results and so now the automation takes care about 30%.

And so now we've gotten down to just about 8 years worth of work and we've been around for 9 years so we can split this work among staffers and it's not anybody's favorite part of the day, but it's also something that's manageable within the organization, plus having some manual review means that we occasionally get to kind of stumble across interesting redactions and sort of fascinating bits that we might not otherwise see such as this wonderful photo of an FBI holiday party that we had a manual a person go through and see.

So what we've found is that sort of as you're building out sort of machine learning process especially because we are a very small team having good fall backs is really important. We're not going to get to a point where we're gonna have 50% of this done or 80% or 100% of this done anytime soon through automation, but having 30% of the work taken off our staffers hands means not only our staffers get to work on other items, but it also means that those users when we can automate it they're getting a much faster response because the machine can run 24 hours a day so even late at night if an agency comes back with information it gets categorized more quickly and so everybody ends up benefiting a lot.

So what did we learn from this process? First of all that data cleaning can take a lot longer than we sort of expected. I think if I had known what we were getting into when we first started I think I might have tried to choose something else a little bit less daunting, but I'm glad we didn't actually know what we were getting into. I think one of the things that's different about us than some of the other great organizations today is that we had no machine learning expertise on staff when we started.

And I think if we had known better than we would have given up, but then I think one of the other things is sort of how key communication is, right? So, I think this is communication with staff and trying to make sure they understand how this machine learning works so they can identify problems. This means communicating with your users so that they understand sort of when things go wrong they have a little more context to understand why that is and they can provide more useful feedback and they're also a lot more sympathetic.

If somebody just sees something weird they don't know they get really concerned, if you give them that context it can be a lot more helpful. And then also it's really important to kind of communicate to your end users who are journalists because journalists are very busy, newsrooms are very busy and essentially newsrooms have been lied to and said hey this technology is a silver bullet, this business models are silver bullet, this is going to save you, so trust us and we're gonna kind of go through with this and I think journalists feel really burn and so they have a very low tolerance for sort of spending a lot of time investing in technology up front if it's not going to work.

And then the other thing we've discovered is that if you make something that was supposed to be bureaucrat proof they're just going to build a better bureaucrat who's going to find some way ways to work around it, so that means it's really important to have a fallback. For example we had agencies just start sending us letters that said "see attached" and they would include an OCR PDF attachment that include the actual information really hard for the computer to kind of guess what might be in that PDF.

So, we did see a lot of success with this though and I think it did help us kind of bring down the amount of sort of manual work by about 66% overall, and so when we merged with document cloud last year we were really interested in sort of what could we do to kind of bring these kind of techniques build this pipeline out for everybody. And so what we did is we launched this crowdsourcing tool that with a drag and drop sort of builder anybody can sort of say hey these are the documents I have, these are the questions I have about them, let me know how to go through them.

And so it's been in a private beta, but we've already had 14,000 responses from over 1,000 users. One thing I hear a lot about is people don't want to see how the sausage is made. People find sort of government documents really boring, turns out there's a lot of nerds who are really really into government documents and so that's been really exciting to see that we're not quite as lonely as we thought we were.

But also this means that we can kind of roll out these tools to newsrooms so that if you have a large set of documents and you need to kind of go through and turn those documents and do data we now have a really nice pipeline even before we add on some of the machine learning work coming later. And it's not just a way to analyze these documents, but it's also a way to kind of engage your readers in sort of building these models, training these models and turning documents to the data.

One thing we found is that readers really like feeling like they're putting in work and investing and understanding this material and having them go through this material with your newsroom is a really great way to do that. And then so, moving forward though we've also found some new techniques that'll mean instead of hundreds of thousands of sort of data entries we can start getting journalists and people who are investigating documents feedback much more quickly about sort of hey we're gonna be able to sort of help you start categorizing this information.

And if you have one hundred thousand letters from an agency to various constituencies and you want to find the letters about a certain topic only giving it a few dozen examples with 20 minutes of work you can start help putting those into different buckets so you can sort of focus your analysis more completely, but one of the things we're most excited with going forward is the ability to sort of help newsrooms kind of share what they've learned and let them build off each other's machine learning projects. So, we want to make it so that you've done a successful investigation, you publish your methodology and then somebody can quickly relaunch that going forward. So, thank you so much.