Data mining and text analytics online communities and applications

Professor Eric Atwell: Hello. This is Eric Atwell. In this lecture, I'm going to tell you a bit about data mining and text analytics online communities and resources. Sort of social media, but specifically for data mining and text analytics professionals. As an AI expert, you certainly will be by the end of your course, you can join these "social media" communities for data mining text analytics professionals to "network", to meet others, to share knowledge and resources.

We don't have to wait till the end of your course because you can join now. And it's particularly useful if you're trying to think of a research project topic and get some ideas by looking at others who have developed text analytics applied research projects. And in your proposal for a project, you might want to think about other related work as part of your background.

First we're going to quickly look at some general social media because they do have AI groups and networks, for example, Facebook, LinkedIn, Quora. And I won't go into a lot of detail because I'm sure you know these yourself. But there are also other communities which were explicitly set up by academics or practitioners in the AI field or the natural language processing field to share knowledge, ask questions, and find resources.

And we're going to look at some examples, KDnuggets, Kaggle, WEKA, ICAME, ACL, SemEval, and the EI-JRC communities. These are just examples, there are others as well. But just to give you a flavor of things you might not realise that you should try looking at.

So the first thing, obviously, is Facebook and other-- what we would call social media. So if you Google Facebook-- I'm not going to go to Facebook to have a look. You can try this yourself. But it's now recently-- the name, the company name has changed to Meta, Meta Platforms, do business as Meta. Formerly known as Facebook, a multinational technology conglomerate based in California. Meta is the parent organization of Facebook, Instagram, WhatsApp, and many other things.

Now one reason for looking at Facebook within the context of text analytics is there are quite a few groups like artificial intelligence and deep learning or corpus linguistics where you can join, ask questions, and interact with users. Trouble is, it's open to all, including Facebook users who think this sounds like an interesting thing to join even though they don't actually know what they're talking about.

However, as well as the standard Facebook, be aware there's also Facebook Research-- let's see if I can click on this and see what happens. So Facebook Research is a huge research lab. I don't want to look at this but you can see they have lots of examples of people who got their jobs at Facebook, publications, research papers. They even invite people to submit proposals.

So if you've got a research proposal and it's good, you can send it off to Facebook and they may give you some money to actually implement it. I'm not suggesting all of the proposals you come up with are going to be that good but they do fund. Facebook basically makes a lot of money and they want to use it to public benefit and part of it is to fund research projects like this. Let's go back to the slide. What next?

We want to look at LinkedIn. LinkedIn is a networking website, a bit like Facebook, for professional contacts. Things like if you're searching for a job or recruiting people. It's become more Facebook-like in that now you can post ideas, news, or whatever. You can follow people and collect followers. You can get notifications sent to you. There's learning and training job adverts and obviously product adverts of various sorts.

I tend to-- when anybody asks me can they link to me I just say yes. So I have accumulated about 6,000 contacts in the AI and natural language processing field. And occasionally I can ask people for advice and so on. Just in case you haven't seen it before, here's LinkedIn.

And this would probably take you directly to my LinkedIn page. Yes. You probably do want to see the details of this. But there's LinkedIn. There's lots of stuff. In particular, my network, jobs, messaging, notifications, work, and so on, and learning. They have some training, not a lot, but it's getting more. So there will be some training courses to do with natural language processing or text analytics.

Next one. You have Quora. You wouldn't think of this as an AI site in particular. Quora is a question and answer website. If you have a question, you can post it, and then other people will answer it. Quora's a place to gain and share knowledge, a platform to ask questions and connect with people who contribute unique insights and quality answers. That means it can be better than just a Google search.

If you have a technical AI or NLP question, like how can I avoid overfitting-- and let's see if I can go to Quora and see if that works. OK, well, there's other questions here. So basically, you type your question.

Let's see. Can I try this? How can I avoid overfitting, question mark, and see what comes up. And it comes up with possible answers from authors. And this is more specific, more directly answering your question than if you just Google search it. And so it's useful for AI and machine learning that's useful for everything else too.

OK. Let's get on to some specific ones in the area of AI and data mining. So one of the longest established websites in this area is KDnuggets, or short for "Knowledge Discovery nuggets." Knowledge discovery is about extracting useful knowledge from data. So it's what in the business school they might call data mining or machine learning. So it's a sort of fancy word for data mining because you're getting-- you're discovering knowledge rather than just getting data and patterns.

And knowledge nugget is a small item or piece for useful information that's often used in business trainers' videos. It's not really a general English idiom. But the KDnuggets website has got news, tutorials, reviews, job adverts. It was originally an academic site and for academics but it's increasingly got lots of AI industry practitioners.

And as you see you first thing you are invited to, "sign up for a newsletter." I have signed up. Now I get sent every week learning posts that people have put in blog pages essentially. How to get certified as a data scientist. There's quite a few quite technical stuff. For example, "2021: A Year Full of Amazing AI Papers- A Review."

So this is a summary of the main findings in AI research papers this past year. That could be good. I would definitely sign up for KDnuggets newsletter. And a lot of it is adverts and stuff you don't want to read but there are nuggets of knowledge in there worth having.

Another really useful website is Kaggle. Originally it was just competitions to do a machine learning, and then pretty soon people realised if you want a data set-- well, companies like Google started contributing big data sets for these competitions and giving big prizes for winning the competition because they realised that lots of enthusiasts would try to tackle their tasks and they might actually get some useful software and results out of this. And it's now been bought up by Google and if you go to the website you can use Kaggle, which is actually Google Cloud resources.

It says, "Kaggle offers a no-setup customizable, Jupyter Notebooks environment. Access free GPUs and a huge repository of community published data and code." And some advertising from Google. And the nice thing also is that the competitions, once they're over, and somebody's won it or not-- not all of them have prizes but they're still fun to-- but afterwards, the data is archived. And very often, the discussions are archived, not just the results and who won, but how they got about how they developed it.

Originally it was just enthusiasts. It's now increasingly taken over by industry sponsors and prizes, but it's still a lot of useful data sets that are available. If you go to Kaggle we'll see top competitions, data sets, code, discussions. That's really useful. If you've got a problem you can take it there.

Let's just look at some of these data sets that they have on offer. Indian coin denomination data, heart disease mortality data set. It's got quite a few medical data sets of various sorts and also sports data sets. Health data sets. There's stuff to do with software trending, food, caffeine content of drinks data set, travel and so on. Lots there.

So one thing you can do, if you're trying to think of a project that you might want to do, then maybe if you've got a Kaggle and find one set that are related to what you're doing, then you can include those in your background. Or the other way around, if you can't think of a project, go to Kaggle, find one that looks interesting and then think of a way of adapting it to your specific company or specific interests. And Kaggle has got lots of models to work from. So that's general enthusiasts.

There's another sort of way of getting enthusiasts who have in common a piece of software. So quite a few platforms for AI. As well as having software you can download, there's also a website behind it. So for example, WEKA. We're going to look at the Waikato Environment for Knowledge Analysis. Originally set up at University of Waikato for their machine learning research group, it's grown and grown.

It's now a hub for WEKA users where you can download software and get links to the textbook, because they've written a textbook to describe it for use in teaching. There's also other tutorials, video tutorials, and so on. Lots of people have contributed dataset to WEKA so you can get those too.

It's just one example because there are other plenty of other software tools that you might want to use and they also have websites to go with them like sketch engine for example. But the Waikato website is much more open and community based because the company hasn't taken over. It's run by the enthusiasts only so it's not really trying to sell you stuff.

If you go there, you'll see it doesn't look as flashy. There's not lots of graphics and pictures but the text is just the basic information about it. You can download the software. There's lots of documentation. Frequently asked questions, how to get further help, there's books and other papers, research projects, which have used WEKA for example. Lots of people involved so if you find that you want a specific question about a specific piece of software, you may even find the person who developed it and ask them directly. So that's WEKA.

Another example of a network. So WEKA is a community of people around a particular piece of software or a particular platform. Another sort of reason for a community, a specialist community, would be people working on a very specific research niche. So ICAME is the International Computer Archive of Modern English.

And this is very specifically for corpus linguistics researchers on English, and in particular for English language teaching, teaching English as a second language or a foreign language. So all around the world, English is an international language, but for many people, it's not their first language and they have to learn it as a second language, and possibly some of you students watching this. And they developed the first English text corpora. The Brown Corpus, the LOB Corpus, The International Corpus of English, and so on.

One of my very first jobs after graduating was to work on the LOB, or Lancaster-Oslo-Bergen Corpus. A million words of British English to compare against the Brown Corpus, a million words of American English. And they hold an ICAME conference every year. It's the 43rd in 2022 so it's been going quite a long time.

And if you look at some of the early ICAME conference photos, you may see me in there. I was at some of the original ones. Not 43 years ago, but nearly. And one of the very useful things they've had for quite a long time is the CORPORA mailing list. So again, if you go to the ICAME website, you'll see it's not particularly flashy, there's no images. This is typical of an academic thing.

We just have links to their conferences and also link to the CORPORA mailing list where you can subscribe. And then once you join the mailing list, then they will send you emails. It tells you what you have to do is send an email to CORPORArequest@uib.no with the line "subscribe" and then and you get subscribed. So it's pretty basic compared to flashy industry web pages but it still works. And that's English corpus linguistics.

A bit broader than that, and in fact, quite a bit broader, is the ACL, The Association for Computational Linguistics. This is a sort of professional body for computational linguistics, natural language processing, and text analytics for academics and also researchers. So people working at Facebook research labs will be in the ACL if they're working on text analytics, as well as University academics.

All the industry research labs like Google, Microsoft, and so on, they present their research at ACL journals and conferences. So the computational linguistics journal publishes quite detailed research papers about particular research projects. They also run conferences and resources and they have an anthology of-- they also have lots of special interest groups. Computation linguistics is still quite a wide area.

I already mentioned, if you remember maybe before, SIGWAC, The Special Interest Group on Web as Corpus, as a sort of people who might use something like sketch engine to collect the corpus and then analyze it. Or SIGSEM, we'll have a look at. They're the Special Interest Group In Semantic Analysis and words, in particular. Let's quickly go to the ACL website if that works. It's actually based in America.

There is some industry interest. They publish a journal but you'll see it looks like an academic web page with lots of text on it. Yes, go back to here and I said yeah, they have a wiki. The wiki contains lots of information. For example, a list of journals in computational linguistics. That's all. There's a list of general AI journals not specific to linguistics. And then journals in cognitive science and psycho linguistics. And there are quite a long list of computational linguistics and natural language processing journals.

So these will publish research papers which you might-- if you're looking for papers on a project in text analytics, this is where you might find them. There's also more specialised journals information retrieval-- that's Google search and things like that, which is a subset of computational linguistics.

Then there's journals on linguistics, which is the theory behind computational linguistics. And these are not particularly computing at all, this is theoretical linguistics. And there's also a couple of journals on machine learning, which are not specific to computational linguistics or computing or text analytics, but have machine learning theory and models which might apply.

And they also run conferences. So they have a web page listing upcoming events, including, for example, the 13th Conference on Language Resources and Evaluation, known as LREC, in Marseilles, and it's happening in June 2022. So you've got the dates.

You also have to know-- if you want to go to the conference, typically, you have to submit a paper. You don't have to, but they ask you to submit a research paper, maybe four to eight pages long, describing your research. And then the deadline for submission is often quite well before the actual workshop or conference. So the deadline on all these conferences, this one's in June 2022, but it's already too late to submit a paper to it because deadline's gone. So that's that.

Oh and all of these conferences and journals and academic go and publish a 4 to 8 page paper in a journal might be longer. All of the papers since the beginning of time, well, since the beginning of the ACL, are collected together in this ACL anthology. So if we go back, you see 1989 and older.

There's a 1974 computational linguistics paper. This is computation linguistics journal from 1974 and you can read the PDFs from this. This is before PDFs existed so what had to be done is to scan the paper versions of these old journals and convert them into PDF that way.

And just out of interest in the ACL anthology, let's try Atwell. I'm going to search for Atwell. See how many papers there are in there. Any search results? Oh, yes. Here's some interesting papers by Eric Atwell. Oh well I'll leave you to enjoy these at leisure So that's the ACL.

And as I said, they have many special interest groups, a bit like ICAME is a specialised interest group for English language corpora and English language teaching. And another special interest group is Web As Corpus. Yet another one is Semantic Analysis. And these special interest groups can host their own conferences and workshops.

So one you might want to look at is SemEval, or Semantic Evaluation. Every year they have papers in the journal-- sorry, in the workshop proceedings, but as well as having anybody who wants to can write a paper on their semantic research. They also have these things called shared tasks. It's essentially a competition, a bit like Kaggle, with a data set and a particular task, like build a classifier for detecting offensive language in tweets, for example.

And they have a data set of tweets and they've had people have gone through and mark the ones which are offensive and ones which are not offensive. But as well as entering the competition, you're supposed to write an academic paper, 4 to 8 pages long, describing your methods, your experiments, and your results.

And then the competition organisers select the best ones, the best paper that is, not necessarily the best results, but the best written papers with novel algorithms or methods. And these are invited at then present their work and there's usually-- the organisers write a sort of summary paper comparing all the different methods and results.

So it's a bit like Cadwell but with academic research papers which you can cite. So again, if you're thinking of a research project, then have a look at SemEval to see if any of the tasks near the recent SemEvals are similar. And then you can cite these papers in your background. Or if you can't think of a research topic, look at the past SemEval competitions. Choose one of the tasks which looks interesting and see if you can modify it a bit to make it more relevant to your particular company or your job, or your interests.

So that's because that way, you can then find-- if you look at the task and the proceedings of the workshop, you can find a range of papers describing different approaches and how good they were, the scores. You can then try the best ones, or some of them, on your task and you can also cite them as relevant background in your research proposal.

And quite a few of these SemEval big data sets have been donated by Google or other companies, so they're really worthwhile data sets. And they've been made available for re-use such as the offensive language identification data set.

So let's have a look at SemEval first. This is the current SemEval 2022 competition. There are 12 tasks on a range of topics. And let's see what they have. For example, CODWOE, Comparing Dictionaries and WOrd Embeddings. We'll find out more what sort of word embeddings are.

And there's also a social task like Patronizing and Condescending Language Detection. That sounds fun. Or Multimedia Automatic Misogyny Identification. Or Intended Sarcasm Detection in English and Arabic. See a lot of these are only for English but some of these competitions are multilingual for other language data sets. It's easy enough to collect English data sets, it maybe a bit harder to collect Arabic data sets or other languages. You can have a look at this yourself later on.

And as I said, some of these data sets have been collected and documented quite in some detail. So here's the Offensive Language Identification Dataset from 2019. So the SemEval 2019 competition proceedings have papers describing all the different contestants, how they did it, how they detected offensive language in the data set. So if you wanted to, you can read how other people did it. It was task six in the summer of 2019.

And there's a paper by the organisers describing how the different competitors fared compared to each other, what algorithms was the best, and which were not so good. And you've got the data set you can download which, for each tweet or each social media post, is marked with offensive or not offensive.

And if it is offensive, in addition, it's got marked is it a generic, horrible slur or is it actually targeted at something. And is it generally nasty or is it actually aimed at a particular target. And then if it's targeted, there's some further information about is it aimed at a group in general or some individual. So there's quite a lot of detailed information in there. That's just one example.

Finally, as another sort of example of a specialised group, so we've seen that ICAME was basically a group of professors of core English corpus linguistics, got together and said, let's set up a website and ICAME and they did and it gradually grew. Another way is to have a group around a piece of software like WEKA.

Another thing is, within ACL, various subgroups keep forming across a very large group of computational linguists and groups within them have formed subgroups or special interest groups.

Another way of doing this is the European Union has been quite keen in fostering research in Europe. So the European Union has a Joint Research Center and they fund groups of researchers to set up communities. So the EU-JRC will fund projects, not so much to do research, but to foster EU collaboration.

So as an example, the HUMAINT, H U M A I N T, project is on human behavior and machine intelligence. It involves a core team of researchers who have-- the grant is essentially do some research but also to set up a community of experts in cognitive science, machine learning, human capital interaction, and economies. European Union wants not just to foster research but actually make money out of it.

So the project organises a series of scientific events. Here's their web page they run a winter school and workshops and other stuff. And because it's European Union there's always a

European Union logo there. And you can find other communities here too. This is just one of them. You click on communities and you can search other communities.

So that's examples. That's not everything here but just to keep you thinking. Hopefully, you're interested in artificial intelligence. You want to broaden your horizons. You're obviously learning but you want to become an expert. So you should join the network. Join these sort of quote, social media, unquote communities for data mining and text analytics for professionals. To talk to each other, to share knowledge, and also to gather knowledge. So if you want some knowledge and resources this is a good source.

Particularly, if you're trying to think of an applied research project proposal idea, then go to some of these to get some ideas. Also, if you're developing a proposal, then you have to have a background section saying how you're building on other people's work. So again, go to these to find other people's work that's related.

And there's this general social media like Facebook or LinkedIn or Quora where you can have specialist groups or you can ask questions to do with text analytics. There's also communities set up by academics and practitioners to share all of these things, to ask questions, and find resources. There's the general AI and machine learning competitions and things like KDnuggets and Kaggle. There's, usually, for any large AI platform like WEKA, there'll be a community around that.

For special areas like English language teaching for ICAME I or semantic analysis like SemEval, there'll be special interest groups that have competitions. They generally have annual conferences and workshops and maybe a journal as an ICAME journal for example. The ACL is the professional organization for text analytics professionals and academics. And it's quite a broad organization. It does cost money to join I should warn you.

Although, you can get free membership to access the website. But they like people to pay so you can get the journal and access to some of the conferences. The European Union, and to a lesser extent the British, EPSRC, the Engineering and Physical Science Research Council, they also try to foster communities and have websites for those communities.

So go out there and join some of these online communities to use them, to get resources, to get knowledge which you can then use in your course works in your projects and so on. Thank you very much for listening. Hope you enjoy.

Oh, and if you want to be one of my LinkedIn or Facebook friends-- I stopped using Facebook, I have to say-- but you certainly can send me a contact to join in LinkedIn and I'll be very happy to LinkedIn to you. OK. Bye for now.

[END]