EDM Webinar

A Hive Mind: New NLP Techniques to Cross-Pollinate Your Cloud and Bloom New Insights

Live Date: Dec 1, 2021

Featuring:
Ian McCarty, Chief Product Officer at eContext.ai
Ben Easley, Technical Evangelist at Datorama

Recording: bit.ly/3oDXdPL

Presentation: bit.ly/3DEktRE

EDM Council Homepage: edmcouncil.org **eContext Homepage:** econtext.ai

WEBINAR Q&A:

Which industry is leading in NLP capabilities?

There's good, interesting stuff being done in certainly marketing and advertising because consumers generate a lot of linguistic data, so definitely want to understand what that is and capture it. There's a lot of processing of written documents in the financial industry because of financial governance and banking. There's just a lot of documents that are created, so there's oftentimes a lot of techniques to process those quickly, and with the speed that financial markets move, efficiency there is super important. We also see it growing in the healthcare space and in the medical space, where again, you have lots of linguistic data from patient records, doctors' notes, and other things that are now quite well established in electronic format. That's a good source to move and apply a lot of these NLP techniques. Other areas are anywhere there are increasing voice interactions, which is happening, in a lot of different places, such as in the car and automotive. In the home, you know everyone probably has a device that they could touch with their hands, they can talk to, and it's trying to understand them, so yes, there are lots of different verticals. At eContext, most of our customers are now focusing a lot on the marketing and advertising use cases, market research, customer service, publishing the right content, and so on.

Is there a software recommendation or approach to building NLP capabilities, such as Python?

Software is language-dependent; I will say that if you're focusing on language, the majority of the high-quality Open Source modeling that you're going to use is going to be in Python; that's really the data science language of choice right now. So if you're looking for skills to build up an engineering team or an ML team, Python is definitely the way to go there. I'll say that there are some other advantages in terms of the way that you might want to store the knowledge graph itself, so a lot of the good graph database technologies out there will allow you to create that structure, not just a hierarchy with parents and child, but using other kinds of edges and predicates between different nodes in the taxonomy. There you're really starting to build more of an ontology which has even greater advantages, so you can look at specific graph database software out there. Here at eContext, we use Neo 4J which is a really powerful extensible graph database to hold and structure our taxonomy, so yes there are a couple of different ways to go, but the taxonomy itself shouldn't be language-dependent, but a lot of the modeling is, as I said, Python heavy.

Is "SO CATCHY" multi-lingual for its NLP capabilities?

That's a really good question. So there are a couple of ways to solve it. One, it certainly depends on who the operator is. If you need the operation or the ingestion of the insights to be used by speakers of different languages, then yes, that's a bigger lift, and you're going to need to have the results be readable by different language speakers. If we're talking about the input documents themselves, there are a couple of ways to do it. You could definitely take multilingual inputs and use those as your training sets so; then you are developing, you may have one consistent hierarchy. But the training data and the way you train the model will be a little bit different across languages. That's one way to do it. The other way to do it is to keep everything the model knows, you know, core language that is, for you, and then anytime a document that you want to process is incoming, you translate that document to whatever language the system understands. That's the way that we chose to do it here at eContext. So, our core engines work in English; that's what they're trained on. But anytime we get incoming documents, we can detect what language they are, and then, if they're in one of the many languages that we support, translate those documents into English to work on them. Then, when we're sending them back out it retrains right back to the source language but aligns the annotations again as you saw on that slide, back to the original source language, so you get those nice classification labels to exist. So there are a couple of ways to do it, and it sort of really depends on where you feel like your skill set lies. Or, if you want to leverage, and there are lots of really good neuro machine translation services out there from many providers, so If you're more interested in using budget that way, then certainly you could build a system in just one language, and then use another provider to do the translation of documents in and out.

Some customer service calls say, "Your calls may be recorded for training purposes." Does that mean the calls could be analyzed using NLP algos?

It could. You know it depends on what each individual customer service provider is doing. Sometimes those recordings just might be literal audio recordings that are, you know, listened to and played back later. But in other cases, yes, those recordings could be transcribed into text using speech to text or voice tech software, and then different processes run on those. It's fairly common, I think, in the industry, to do some basic NLP processing, even if it's just speaker separation or sentiment classification. There are some other tools out there that I'm not quite sure I'd call them exactly NLP, but they will work more on audio rather than text. So there are some cases where call recordings can be analyzed for the sound itself, the tone. Whether that's the rate of speech or the pitch, the timbre, etc., to derive some insights into what the customer might be feeling or how that call is going. So there are different ways you could use a recording. For us, definitely, getting that into text gives you, I think, a lot more capabilities and different avenues to pursue.

Has this type of tool been used to help build business-term glossaries, and where terms are used?

Yes, absolutely. If you're building a business glossary, then you know you're on the path towards this kind of thing, and that's fantastic. A good robust business glossary would really just be one step away from a good taxonomy if you start to add in those connections and those linkages of where one term is a type of another, a more specific example of another, right, where those things exist in relationships to each other. So yes, if you've already started with that, starting to add those connections can get you to that next level, and then really help as you move to the analytical stage of being able to do the zooming in and zooming out, and the aggregations that we talked about.

NLP is a challenge with short text. Can you categorize text that is two to ten words?

Yes, absolutely, and that's going back to why we recommend training on those really short pieces of data, those quarks. As a little bit of history, at eContext, we originally started as a technology that was working strictly with search keyword data. We were working on behalf of a Metasearch platform that we used to run, so all the data that we originally looked at was search queries which are one to normally fewer than ten words. So it was, at the time, a necessity, but now we realize it's a benefit, where if you do the hard work of understanding the shortest pieces of content, and you have those as letters, instead of words, to type with, then you can write something that's much more eloquent. So, yes, a lot of NLP does struggle with that, and usually, that's just because of the effort to create labels at those very short text lengths. It's much easier, and you can shortcut your way to the systems, if you just know that this document is all about finance, or that document's all about automotive, and then we'll try and find patterns in there. But, as we know, that can be problematic if you want to try and use that model for something that isn't a full document.

How many years have you been building and developing this Taxonomy? What types of data have you been using and training on?

That's a good question. So, we've been around for many years. We were established in 2009, so we've been doing this for a long time, and we have seen data evolve over a long time, which is certainly valid if you can get in your training site, and you can get data over a large amount of time, and that can be really valuable. So, because we know that our clients are coming from a lot of different verticals, and they have a lot of different needs or types of documents, we purposely train on quite diverse data sets. That includes search keywords, as I said, and a lot of short, user-generated content like social media and forum posts. But then, also, longer documents, news and articles, and websites. There are a lot of great resources out there, the common crawl of all the websites on the Internet, which we processed, or in a lot of places, will go and use Wikipedia data, also, because that's pretty available. But, in our view, it's best to pull from all sources so that you are as best equipped as possible to handle these different channels and these different verticals.

Is your taxonomy curated by humans, or do you use machine learning and AI?

Both! Humans built the initial concept architecture and crafted business rules to categorize input to any node in the taxonomy. From there, those same processes were used to rapidly label billions of pieces of text and then train a machine-learned model that can also predict the same classes from the hierarchy. Today, those systems run in a hybrid mode, supporting each other to achieve both high-precision and high-recall.

How prevalent is high-frequency algo trading-based NLP of social media feeds today?

We can't say for certain how often NLP is used to monitor social media feeds as a source of insight for algorithmic trading, but it is definitely a rich use case. My guess would be that it is not used as much as those financial stakeholders would like, probably because there is a dearth of NLP systems that can extract enough granularity and structure from social media to properly align with trading decisions.

Sometimes the same term has different meanings in different areas of the business. How do you segregate the labeling for such terms?

That's a really good example. What you'll need to do there is start to expand out the size of some of those quarks that we talked about. So you're going to want to find good examples where that same word or short phrase is embedded in other content that gives it that greater context and that greater meaning. When you're looking at applying labels to things, or building up that training set, make sure that you're pulling out a little bit more of the surrounding language, and letting the machine learning models that you leverage, giving them the best chance to find those patterns, oftentimes patterns that aren't really visible to us as people. So, you're going to want to make sure that you have good examples, and also that you have enough classes, so that you can represent them, all those different uses, because, again, you know computers just do what we told them to do. If you tell the machine that you know this term always means this, and that's the only sample you give it, then even when a different use or appearance occurs, it’s still going to label it that way, because it has no other examples. That's another reason why we want to build this wide knowledge structure, so that you can capture and well represent in your data all the different ways that that term can be used or can have meaning.