Market Abuse Surveillance TechSprint (July 2024) video Transcript.

Team 5. SteelEye

Delegate 1

Hi everyone.

We are SteelEye, I'm Ivy Lame. This is Hisham and this is Louis from Brainpool, our collaborator in this we are. So we are collaborating in terms of our solution for this text Sprint and our solution is an advanced detection for layering cross product layering.

So who are we? We are Still Eye, the industry's pioneering integrated trade and surveillance platform. So we know that in the recent decades there's a big, big explosion of data and we do realise that in order for our clients to meet their regulatory obligations, we need our platform to be data centric. So at Still Eye, our platform is data-driven and along with the with leveraging the latest in AI, we give our clients a complete holistic oversight of all trades and comms activities.

In this TechSprint, what we focus on is the improvement of our current way of detecting market abuse. So right now, we have what we, what we call the traditional method. So within this TechSprint in that small space of time of 2 1/2 months, what we did is to create a brand new novel solution.

So our solution is focusing on identification of the more complex difficult to identify market abuse. So we do know that one of the issues is the fact that relatedness in terms of cross product especially increases the number of pathways where bad actors can manipulate the trade, the markets itself. So to overcome the issue, we develop this novel solution and it comes with, as you can see, three phases.

So the first phase is for us to correlate the data set that we're given. So all the relevant data set, we correlate all in to connect all the dots in preparation for the next phase.

In the next phase is what we call anomaly detection. That is where we detect. Well, the goal of that phase is for us to eliminate the true negatives, pull out or philtre out what we call the positive or the potential positive for market abuse, these subset of orders or trades or transactions.

Then we pass to our third and final phase, which we call MA verification. In this phase, we use an LLM with all its free train knowledge and what have you to go through each of the subsets of orders and determine whether these are true positive of market abuse abuse, where you can then give it to our end user for them to do further investigation with this high level overview.

I'm going to pass you to my colleague Louis, who will go through this solution in detail and give you guys our finding.

Delegate 2

Thank you very much, Ivy. Yes, so I've spent the last two months working with the text print data. There's a lot of it. And so I've sort of honed it into a small time window.

I've looked at a particular day in particular and I restructured the data into a vector database. And the reason I chose a vector database to start is that you can build relationships into vector databases such as intra trader relationships, which might help detect patterns of collusion, as well as instrument building up instrument relationships which can be built up by correlation matrices. You know, you can look at industry sectors, you can even use LLMS on the descriptions of them to find correlations if you wanted. And it's from there that you can build out cross product detection methods.

The other benefit to using vector databases is that it integrates very naturally with any natural language data, data sets. So news data, comms data, anything like that. Vector databases, the sort of LLM comes sort of out the box with querying it comes part of the querying of the data which is really quite useful and powerful.

So the first thing I did was I, I started with the full, full orders of something like 63,000,000 for the day and I've reduced it, filtered it down to some pre processing down to 15,000,000 which got read into the vector database. And then from there we wanted to reduce it down.

I developed a novel algorithm based on entropy. So entropy is a measure of disorder. And so if you can look at a series of orders that are in quick succession with one another, you can say if they're all sell orders, that's low entropy. If they're all buy orders, that's low entropy. But if there's a mix mixture of the two, the entropy goes up. So we actually want the entropy to be high for, for what we would say is genuine trading. And if we see low, if we see very low entropy with an A counter order or a counter fill, that's on the other side of the order book.

That is an indicator that we might be dealing with dealing with layering. And it integrates very easily with cross product because you just simply philtre on

what the set of instruments as opposed to the single instrument based off that correlation matrix. So from there, from that algorithm, we found that we could we could reduce it down to 29 viable candidates simulation. So it comes up with a score.

So it ranks them all and found that a score of 40 from simulated data was like, that's, that's decent, decent layering. And what's interesting is I found that on this given day, we found that we got 29 of these anomalous orders amongst 4 traders. Interestingly, one of those traders seemed to be responsible for 21 out of those 29. And the scores were off the chart like three or four times bigger than the simulated score. So this is very indicative. And I, I said at the start of this project, the success of this project for me would be, can we actually find people with the data that we've been given? And I believe that there is viable cause to investigate certain, certain, certain participants in this data.

Of course, we can't say for certain. There's an investigation process and that's where the LLM steps in. So with those, we feed those orders into an LLM and you can, this gives you explain ability out the box. You can even ask the LLM what explainability you want. You can say layering doesn't happen on the order of hours or minutes. It happens on microseconds. You tell the LLM, here's a series of orders.

These are the definitions I have of, of layering and spoofing. So tell me what you're finding here. And what's interesting is the LLM correlated very well with my score. It, it correlated very well with my score, a score of 10 to 20.

I actually used two LLMS and there was disagreement between the two of them. But as that my, my entropy score went up the, the, the ranks, the LLMS were verifying what you can verify humanly by I. And so you get the explain ability out out-of-the-box for the LLMS, which I found to be very, very, very useful. And it verifies by I because you you could manually go through these orders, but it's a tedious process and LLMS can do that for you. So I found that to be really quite a powerful way of filtering even further down some of those results. Yeah.

Thank you very much. Back to Ivy.

Delegate 2

Thank you, Louis. That's great.

So moving forward, this is what next steps you can do in order to buffer the solution into something that is more well-rounded. There is actually a fifth one, which is like something that we hope we could do, but we couldn't.

And the fifth one is for us to fine tune the LLM to in in order for it to specifically or to enhance its ability to detect, enhance its ability to detect market abuse And very quickly the force are expanded and type of market manipulation to be detected and in more different type of data sets like ECOMS add in more different type of instruments and increase order book depth.

Thank you very much.