Bond Case Briefs

Municipal Finance Law Since 1971

AI And The Municipal Bond Market: Oceans Of Data

The municipal bond market is often described as complex, opaque, and fragmented with multiple sectors, submarkets, and tens of thousands of issuers. Documents and financial statements are mostly in PDF form, not digital. Financial statements lack a consistent taxonomy. All of which contribute to the self-perpetuating myth of that the market is difficult to nearly impossible to neatly organize and categorize.

Except now, with AI technology, this myth is about to be busted. Finally.

In Plain English

In developing English language version of ChatGPT, the engineers had to code in not only all of the spelling, grammar rules (and the exceptions), and definitions of words, but also the various meaning of words both independent of and in conjunction with each other. In all, they had some 400 billion words—individual, combined, and sometimes repeating—to wrestle with. This took up around 570 gigabytes, running through a variety of complex learning algorithms.

Key to ChatGPT developers' success was taking words, which are unstructured data, and making them into structured data. Basically, unstructured data—words, photos, sounds—come with no inherent numbers or tags a computer can read or understand. Structured data, like numbers in spreadsheets with labelled rows and columns, comes tagged and organized, readily computer legible.

What makes AI technology so formidable is the well-defined data architecture frameworks that can organize, tag, and categorize whatever is presented as unstructured. This is generally referred to as tokenization, where a word or part of a word is assigned a number.

It's not limited to words. Take anything unstructured, apply a number and a tag to it, and it can be transformed into data.

Anything. That includes all the unstructured data in muniland.

Clear Sailing on Oceans of Data

When it comes to data, the municipal bond market has oceans of it. Trading levels, offering documents, financial statements, deal structures, yield curves, professional publications. Name just about any issuer or bond data and it exists.

And, contrary to what some in the market have asserted, from an AI standpoint, the data is very, very good.

Here's why. For one, the reference data is generally standard for bonds. No matter the sector or issuer, there is a coupon, maturity, and so forth. There is a low need for unique identifiers. In most cases, a rose is a rose is a rose is a rose. Thanks or apologies to Gertrude Stein.

Second, as a whole, the market has developed a standard professional vocabulary. Many of these terms are fairly consistent from bond issue to bond issue and from sector to sector. It may be a rare instance where we can actually thank bond attorney's for repetitive boilerplate disclosure language and presentation formats. By codifying the market's linguistic traditions, counsel may have actually paved the way for AI to be applied easier.

The irony that attorneys inadvertently contributed to drafting this law of unintended consequences is not lost.

As for the data overall, one MIT GOV/LAB data researcher noted in a recent publication where he applied various AI analyses on over 4 million bonds in 445,000 issues, "municipal bonds have a high degree of transparency, with large, consistent and easily available datasets stretching back many decades." This may be the first time "municipal bonds" and "transparency" have appeared in the same sentence without the word "lack" between them.

Show Me The (Unstructured) Data

It's not hard to find data in the municipal bond market. It's everywhere. Numerous data aggregators and providers proliferate, including Merritt Research Services, DPC DATA, Bloomberg, Mergent, LSEG, S&P Global Market Intelligence—and others.

But arguably the world's most comprehensive and publicly accessible municipal market database can be found in the Municipal Securities Rulemaking Board's Electronic Municipal Market Access platform, or EMMA for short. As the principal regulator of the municipal securities market and part of its legislated mandate, the MSRB is the repository for all municipal securities disclosure. If it pertains to a publicly offered municipal security, from Official Statements to Annual Comprehensive Financial Reports (ACFR) to rating changes to bond trades and everything in between, it has to be filed with the MSRB. Municipal trading data and disclosure documents associated with municipal bond issues are available at no charge on EMMA or on a subscription basis in real-time for a fee. In fact, many of the market's fee-for-data providers draw from EMMA.

Structured data, such as on trades, is straightforward enough to download and, after some massaging, readily machine readable for AI analysis.

But it's the enormous volume of rich, unstructured data in the MSRB files that proves more challenging. Start with the Official Statement, referred to as an "OS" by market professionals. The OS is the equivalent of a prospectus for a stock offering. (Note: eventually, the OS will be fully digitized and structured, but we're not there just yet.)

These offering documents have everything an investor could want about a bond issue—final pricing structure, the issuer, the borrower, rating, use of proceeds, operating information, tax opinion. The list goes on. If it pertains to the bond issue, it's in there.

Measuring Up

To take a digital measure of these documents, we first draw from the cornucopia of information in the MSRB Factbooks that a fair estimate of around 185,000 Official Statements were filed from 2009 to 2023.

The MSRB did some analysis on a small portion of Official Statements submitted for a bond issuance. It was found the average OS ran about 150 pages. Multiplying the average number of pages by the number of OS filed over that period and the result comes in at 27,750,000 pages, give or take.

Let’s convert those pages into digits. In digital terms, this sentence has about 52 bytes. Each character, which includes spaces and commas and periods, is roughly one byte. A kilobyte has 1024 bytes, somewhat short of a half a page of text, so a full page is around 2.5 kilobytes. A megabyte is 1024 kilobytes and one gigabyte has 1024 megabytes. By some estimates, there are around 178,000,000 words in 1 gigabyte.

A few jabs at the calculator show those 27,750,000 pages weigh in at some 66 gigabytes of data.

That’s a fair amount of data, but it leaves out the continuing disclosure bond issuers have to file with the MSRB (well, really the issuer’s bond underwriter but that’s another article). Continuing disclosure includes information such as financial statements, redemptions, notice of default, and any other filings required under the law. It adds a lot of data. The OS gets filed once, but the continuing disclosure lasts as long as the bond is outstanding. Some bond issues have 30 year maturities.

Using the MSRB’s powerful EMMALabs Disclosure Search Tool, an exceptional technology to pull up numbers from both primary market documents and continuing disclosures, the Lab currently has more than 36 million pages extracted from over 860,000 indexed documents from 2019 to the present, with the average document being 42.6 pages.

A few more jabs at our calculator under this methodology to disclosure documents calculates to around 85 gigabytes of data.

Drawing the Line

The MSRB’s line stops there. However, if you go one step further and combine the data from the two methods and time frames, do some back-of-the-excel-spreadsheet calculations and extrapolations, you could estimate the all the OS and disclosure data in those files totals up to a very rough 244 gigabytes. While it is admittedly far from perfect, even if you’re off by 10% or 20% on either side, you’re likely still in the ballpark.

Myth Busted

Now, remember those 570 gigabytes the ChatGPT engineers needed to organize the English language?

The municipal bond market’s vast trove of unstructured data isn’t even remotely as complex. The digital amount of unstructured data in those OS and disclosures and financial statements maybe come to half of what ChatGPT required. Moreover, it is more readily standardized, organized, categorized than English.

From Opacity to Transparency

Training ChatGPT models on municipal market language will likely take far less time to develop a MuniGPT. There is no question it’s underlying deep learning technology can be applied to the words quietly sitting in those OS and disclosures and financial statements. The benefits to investors and issuers alike from the fount of information that could be released by machine and deep learning analysis to this data is staggering to consider.

But what AI will do is remove the complexity and opacity to de-myth-tify municipal bonds and make them all a bit more generic.

That would not be a bad thing.

Forbes

by Barnet Sherman

May 22, 2024

My genuine appreciation to the Municipal Securities Rulemaking Board for their work in researching and providing data for part of this article.

Copyright © 2026 Bond Case Briefs | bondcasebriefs.com