PROJECT BANKTRACK

In collaboration with Inclusive Development International (IDI) and BankTrack, the team at the Data Science Clinic worked to track private institutional funding of "bad actors" — projects or initiatives that infringe upon human and environmental rights. For example, JP Morgan Chase, a large international bank, has been loaning out cash to a food processing company called JBS. JBS uses these funds to process beef in the Amazon rainforest. Given the size of its operations, it is implicated as a major contributor to the deforestation of the Amazon. In this case, JBS is the bad actor, and JP Morgan's role as a financier of JBS's operations is not widespread public knowledge. Moreover, these loan agreements are complex and detailed, consisting of varying terms and multiple entities. Although this information resides in public SEC financial disclosures, it is presented as unstructured prose. Thus, extracting relevant loan information from these dense and complex documents is an extremely manual and time-consuming task.

The team worked to automate this task by building an NLP pipeline. This pipeline can take any SEC 8-K form and extract information about loans and bonds, if present. This extracted information can then be stored in an easily searchable online database. This database is central to IDI's and BankTrack's long-term vision for improving financial transparency and accountability.

The NLP pipeline consists of three steps. First, the **Few-Shot Model** identifies whether individual 8-K documents contained loan or bond information. Next, the **Semantic Similarity Model** evaluates individual sentences or passages to identify portions of the document that discuss loans or bonds. Finally, the **Question-Answering (QA) Model** evaluates the relevant portions of the document to extract key information such as the name of borrower, name of lender, loan amount, signature date, transaction type, loan maturity date, *etc*. The results from the three models were promising:

- The Few-Shot model classified documents as "loans," "bonds," or "neither" with **88.3%** accuracy.
- The Semantic Similarity model correctly identified **93.8%** of cases containing loan/bond information.
- The QA model extracted information about the loan amount and signature date with **89%** and **84%** accuracy, respectively, but needs further fine-tuning for information about the name of borrower and transaction type.

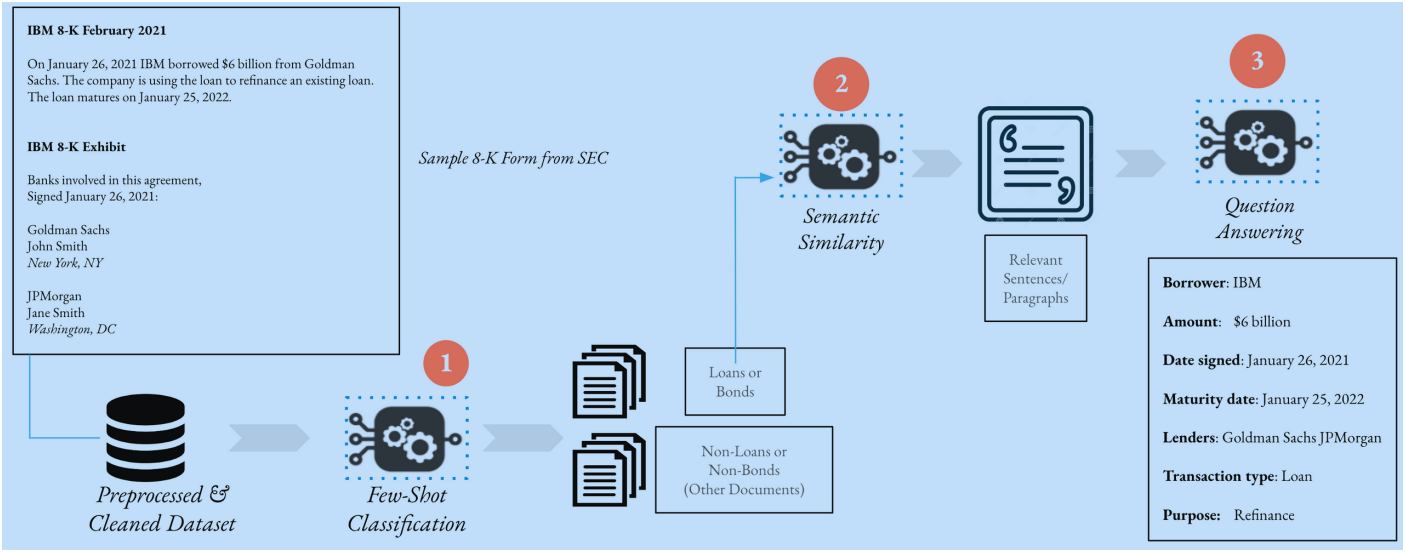

Figure 1: NLP Pipeline Setup