Understanding Financial Reports using Natural Language Processing

Project Plan

Tarun Sudhams, Saripalli Varun Vamsi

30th September, 2018

1 Project Background

This project investigates how mutual funds leverage credit derivative by studying their routine filings to the U.S. Securities and Exchange Commission. Credit derivatives are used to transfer credit risk related to an underlying entity from one party to another without transferring the actual underlying entity.

Instead of studying all credit derivatives, we focus on Credit Default Swap (CDS), one of the popular credit derivatives that were considered the culprit of the 2007-2008 financial crisis. A credit default swap is a particular type of swap designed to transfer the credit exposure of fixed income products between two or more parties. In a credit default swap, the buyer of the swap makes payments to the swaps seller up until the maturity date of a contract. In return, the seller agrees that, in the event that the debt issuer defaults or experiences another credit event, the seller will pay the buyer the securitys premium as well as all interest payments that would have been paid between that time and the securitys maturity date.

CDS is traded over-the-counter, thus there exists little public information on its trading activities for the outside investors. However, such information is valuable. CDS is designed as a hedging tool that the buyers use to protect themselves from potential default events of the reference entity. Besides, it is also used for speculation and liquidity management especially during a crisis.

Before SEC has requested more frequent and detailed fund holdings reporting at the end of 2016, mutual funds filed the forms in discrepant formats. This made it extremely difficult to effectively extract information from the reports for carrying out further analysis. There exist some previous studies that explored how mutual funds have made use of CDS (Adam and Guttler, 2015, Jiang and Zhu, 2016), but only examined a fraction of institutions over a short period of time. In this project, we aim to extract as much CDS-related information as possible from all the filings available to date to enable more thorough downstream analysis. This information appears not only in the form of charts but also in words, thus Natural Language Processing (NLP) is the key.

2 Objectives

The followings are the objectives of this project:

- Understand how the portfolio holdings of CDS were reported on Form N-CSR(S) and Form N-Q.
- Identify the proportion of mutual funds using CDS.
- Among the filings with CDS information, identify:
  - the proportion of filings that are easy to extract. (e.g., well tabulated with pure numbers)
  - the proportion of filings that could be extracted using NLP. (e.g., holdings are described in words)
  - the proportion of filings that cannot be effectively extracted.
- Extract the characteristics of the CDS holdings:
  - Identifier of the mutual fund (e.g., CIK, name, etc.)
  - Metadata (e.g., reporting period, etc.)
  - The particulars of the CDS holdings:

        - Reference entity
        - The direction of trade (buy/sell)
        - Notional amount
        - Currency
        - Contract premium
        - Counterparty

- After extracting the CDS information, some downstream analysis should be explored:
    - Time-series analysis of the mutual funds utilization of CDS during 2004 to 2016. (normal times $\rightarrow pre - crisis \rightarrow crisis \rightarrow post - crisis \rightarrow regulated \rightarrow callforreviving$)
    - Explore the pattern of using index CDS, sovereign CDS, and single-name corporate CDS.
    - Explore whether the funds using more complex sentences describing their CDS positions make more money than those using concise tabulated numbers.

3 Methodologies

We first crawl the US Securities and Exchange Commission website to get the reports of all the Mutual Funds from 2003- 2016. We do this using Python scripts with web crawling frameworks such as Beautiful Soup and Scrapy with Selenium drivers. We also take some open source crawlers such as SEC Edgar crawler which can be used to download all the companys periodic reports filings and forms from the Edgar database.

We then use Perl and regular expressions to extract data from the reports downloaded. Perl is useful and better than Python in this because:

1. It borrows its syntax from C and other UNIX commands like awk, sed etc. due to which it has powerful and built in regex support without importing any third-party modules. While doing string and regex operations like replacement, matching, Python is outperformed by Perl since Perl can achieve the same result as Python in a single

line. Exception handling and many file I/Os are also done much faster on Perl.

2. Perl can handle OS operations using built-in functions. On the other hand Python has third-party libraries for both the operations i.e. re for regex , sys for os operations which need to be ensured before doing such operations.

We will be using Natural Language Processing with Recurrent Neural Networks to process the complex sentences in the extracted data. The best way to represent text meaning is through words. This is why we use word embeddings which lets us represent words in the form of vectors. The aim of this is that words in a similar context or similar words are close together in the vector space while other words end up far away from each other. We will be using word2vec and glove techniques to determine these embeddings.

We represent each word through a unique index mapping. We will have to learn the representation of V distinct words in the form of some D dimensional vector each. Since our vocabulary size is V and we need a D dimensional representation of each word, we need a projection layer consisting of a V*D matrix. Both Glove and word2vec are just 2 different methods to generate the projection layer. Word Embeddings are an unsupervised technique and hence are a very useful starting point for implementing neural networks using Natural Language Processing.

Once we start embedding words, it is useful to use Recurrent Neural Networks because it makes use of sequential information. RNNs have a memory which is used to capture information about what has been done so far. We will be further conducting time series analysis using LSTM-RNN to forecast the buying/selling activities of Credit Default Swaps by Mutual Funds.

4 Schedule and Milestones

- September 2018:
  - Research on Credit Default Swap.
  - Literature Review: Ongoing Research Papers and key financial terms.
  - Familiarization with Perl and RegEx.
  - **Phase 1 Deliverable on 30th September, 2018**
- October - December 2018:
  - Crawling Data from SEC EDGAR
  - Extracting the data from the reports crawled
  - Testing Natural Language Processing Methods
- January - February 2018:
  - Preeliminary Implementation
  - Detailed Interim Report
  - Initial Presentation Deck
  - Exploring Time-Series Analysis methods and possible implementation
  - **Phase 2 Deliverable between 7-11th January, 2018**
- March - April 2018:
  - Structuring the findings of Time-Series Analysis
  - Use data visualization tools to visualize the data analyzed
  - Testing and Finalizing Implementation
  - Drafting Final Report
  - Final Presentation
  - Project Exhibhition
  - **Phase 3 Deliverable on14th April, 2019**
- **Final Presentation between 15-19th April, 2019**

References

[1] Adam, Tim Guettler, Andre *Pitfalls and perils of financial innovation: The use of CDS by corporate bond funds," Journal of Banking Finance, Elsevier, vol. 55(C), pages 204-214..* 2015.

[2] Investopedia: Credit Default Swaps *https://www.investopedia.com/terms/c/creditdefau*

[3] DEEP LEARNING FOR NATURAL LANGUAGE PROCESSING : PART 2 - RNN. *Deep Learning for NLP - RNN, 9 Nov. 2017, techblog.gumgum.com/articles/deep-learning-for-natural-language-processing-part-2-rnns.*

[4] Deep Learning for NLP : Word Embedding. *Deep Learning for NLP - RNN, 4 Jan. 2018, techblog.gumgum.com/articles/deep-learning-for-natural-language-processing-part-1-word-embeddings.*

[5] Kompella, Ravindra, Using LSTMs to Forecast Time-Series Towards Data Science. *Towards Data Science, Towards Data Science, 17 Jan. 2018, towardsdatascience.com/using-lstms-to-forecast-time-series-4ab688386b1f.*

[6] business-science.io., Time Series Analysis: KERAS LSTM Deep Learning - Part 1. *Business Science, 18 Apr. 2018, www.business-science.io/timeseries-analysis/2018/04/18/keras-lstm-sunspots-time-series-prediction.html*

[7] Rahulrrixe. , Rahulrrixe/Sec-Edgar. *GitHub, 25 July 2017, github.com/rahulrrixe/sec-edgar*