Natural Language Processing Report

UNDERSTANDING FINANCIAL REPORTS USING NATURAL LANGUAGE PROCESSING

April 14, 2019

Supervisor: **Dr. RuiBang Luo**

Tarun Sudhams (3035253876)
Varun Vamsi Saripalli (3035242229)

Written by Tarun Sudhams

Abstract

Credit Derivatives are considered excellent tools to hedge the credit risk of an underlying entity from one party to another without actually transferring the ownership of the entity. One such hedging tool is called Credit Default Swaps (CDS), which are often known to be responsible for the 2007-2008 financial crisis. Upon further investigation, it was found that the lack of regulation and information on how CDS works were the main culprits behind the crisis. Post the crisis, the United States Securities and Exchange Commission(SEC) has requested for frequent and more detailed reporting from the mutual funds about their current position on these derivatives. Given the lack of strict format for these reports, it becomes extremely difficult to extract information from these reports and conduct in-depth analysis on how the mutual funds leverage credit derivatives and in particular, CDS.

This project aims at consolidating all the mutual fund holding reports on Credit Default Swap positions from 2004 - 2017 and aggregate into a structured database. Owing to the nature of these reports, a methodology to extract information from both structured and unstructured types of reporting had to be devised. Rule-based and Natural Language Processing techniques allowed us to extract structured as well as unstructured information which often hidden in the semantics of a sentence. Other than extracting and aggregating information, this project makes significant contributions in the finance specific domain by using Conditional Random Field models to extract information instead of just using traditional rule-based approaches.

Finally, upon the successful aggregation of all the credit default swap mentions, we conducted downstream analysis to further understand and answer different questions surrounding the usage and filings of credit default swaps before, during and after the financial crisis in 2008.

Acknowledgement

We would like to thank our supervisor Dr. RuiBang Luo for his continuous guidance throughout the project by providing his expert opinion on various problems that we faced. We are also grateful to Dr. Ricky Ma to help us setup and troubleshoot the development environment.

Finally, we want to thank our co-supervisor Jun Yu Wang, who helped us with complex finance terminologies and supported us in creating a motivation for this project from a finance perspective.

Contents

List of Tables

List of Figures

12. Trend of CDS reporting by Single Name CDS

13. Trend of CDS reporting by Sovereign CDS

1 Introduction

The lack of a structured database of financial reports makes it difficult for Credit Default Swap related research studies to conduct a much more comprehensive and quantitative analysis and also result in inaccurate case studies when it comes to critical topics like predicting the next financial crisis. Therefore, a structured and well maintained database can help future research papers to analyze CDS and retrieve new and exciting information from it.

1.1 Background

Credit Derivatives have a wide range of products and we will be studying a class of credit derivatives called Credit Default Swaps(CDS). Credit Default Swaps have a reference entity linked to them which are generally governments or corporations. The buyer has a credit asset with the reference entity and buys a CDS from the seller to insure himself against a default in the payment by the reference entity. It is thus used as a hedging tool to reduce the risk associated with a credit asset [6]. The buyer makes periodic payments to the seller till the date of the maturity of the contract and this constitutes the spread of the CDS. In the event of a credit default, the seller has to pay the buyer of the CDS the face value of the credit asset and all the interest payments that the buyer would have earned between that time till the date of the maturity of the asset.

Credit default swaps are traded over the counter and hence there isnt much information available on it. The forms filed by the Mutual Funds regarding their CDS activities were in an unorganized manner before SEC had requested for more frequent and detailed fund holdings at the end of 2016.

Counterparty	Reference Entity/Obligation	Buy/Sell Protection	(Pay)/Receive Fixed Rate	Expiration Date	Notional Amount	Unrealized Appreciation/ (Depreciation)
Citibank N.A.	Georgia Pacific 8.125% due 5/15/2011	Buy	(3.55)%	12/20/10	$ 1,030,000	$ (42,489)
UBS AG	CDX.NA.IG.6	Buy	(0.40)%	6/20/11	$ 55,217,800	(74,344)
Morgan Stanley	CDX.NA.HY.6	Buy	(3.45)%	6/20/11	$ 5,501,000	11,923
Morgan Stanley	CDX.NA.HY.6	Buy	(3.45)%	6/20/11	$ 5,501,000	(25,759)
UBS AG	CDX.NA.HY.6	Buy	(3.45)%	6/20/11	$ 5,501,000	(97,046)
Morgan Stanley	CDX.NA.HY.6	Buy	(3.45)%	6/20/11	$ 5,501,000	(21,794)
Citibank N.A.	Windstream 8.125% due 8/1/2013	Buy	(1.60)%	9/20/11	$ 2,885,000	(15,843)
UBS AG	Windstream 8.125% due 8/1/2013	Buy	(1.63)%	9/20/11	$ 2,750,500	(16,547)
						$ (281,899)

Figure 1: Credit Default Swap reporting in a N-CSR report

Credit Default Swaps on Corporate Issues—Buy Protection

Reference Entity	Counterparty	Implied Credit Spread at December 31, 2016[(k)]	Industry	Fixed Deal Pay Rate	Maturity	Notional[(l)]	Fair Value[(d)]	Unamortized Premiums Paid	Unrealized Appreciation (Depreciation)
Hovnanian Enterprises, Inc.	JPMorgan Chase Bank, N.A.	13.2%	Consumer Durables & Apparel	5.0%	9/20/19	$(5,000)	$859	$796	$63
Hovnanian Enterprises, Inc.	JPMorgan Chase Bank, N.A.	13.7%	Consumer Durables & Apparel	5.0%	12/20/19	(2,000)	388	371	17

Figure 2: Credit Default Swap reporting in a N-Q report

The length of each report could span hundreds of pages making it difficult and tedious to employ humans to gather information. This resulted in it being extremely difficult to get relevant information from these reports to carry out further analysis. Thus, information regarding CDS is extremely valuable as it would provide transparency and can be used to set appropriate capital requirements for financial institutions trading CDS. There have been a few previous studies exploring the usage of CDS by Mutual Funds[1],[13] but these reports examined only a small number of the institutions over a short period of time. Hence, we choose to comprehensively examine all the reports from 2004-2017 and this makes the results of our project extremely valuable for further research.

1.2 Objective

The objective of this project is to aggregate a structured database of credit default swap reportings from 2004-2017. This paper proposes a natural language processing technique called sequence labelling using Conditional Random Field to highlight key CDS information in unstructured sentences. Then, the project aims to answer based on the CDS reporting and package the database into a web application to power future research studies on the effects of CDS.

1.3 Scope

In the area of Named Entity Recognition, the sequence labeling of sentences can be conducted using two methods, namely Bidirectional Long Short Term Memory - Conditional Random Field or a simple Conditional Random Field.

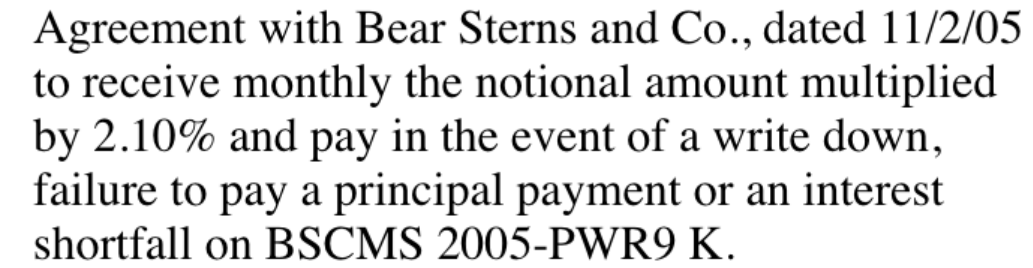

Figure 3: Unstructured sentence reporting of CDS in a report

However, when we successfully extracted raw(data before extracting the information) structured and unstructured CDS information, we could not find enough unstructured reporting of CDS to implement BiLSTM-CRF with high precision and recall. Therefore, this paper will focus on the implementation of a CRF model to extract information from unstructured sentences.

Furthermore, unlike the similar study conducted earlier (Wei and Zhu, 2016) which only took into account of reports filed between 2007 — 2011, this project aims at analyzing all the data publicly available which is from 2004 — 2017 and use it to conduct a significant downstream analysis and draw insights from it.

1.4 Deliverables

The complete implementation of the project is available on here . The project has a few major deliverables which have been outlined below:

1. **Text Annotation Tool** - A Django based web application developed to produce custom datasets for sequence labeling projects and allows for collaboration between users to tag sentences with custom entities.

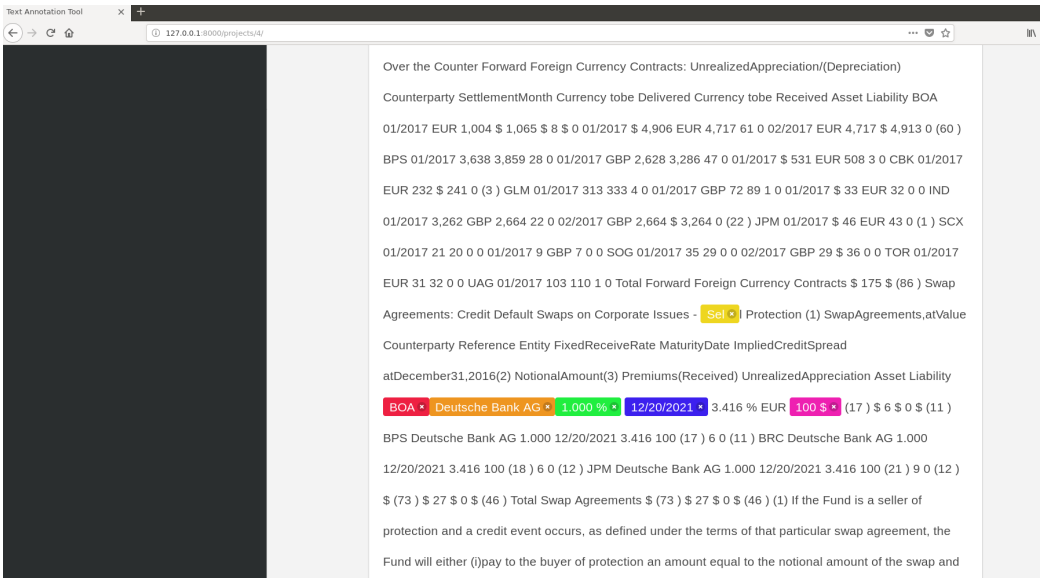

Figure 4: Tagging sentences in Text Annotation Tool

2. **Credit Default Swap Reporting Dataset** - A 16,813 rows dataset containing CDS reporting from 2004-2017 with categorical variables like Reference Entity, Reference Obligation, Reference Entity, Notional Amount, Expiration Date, etc.

	CIK	Reporting Type	Reporting Year	Counterparty	Notional Amount	Reference Entity/Obligation	Expiration Date	Appreciation/Depreciation	Upfront Payments Paid/Received	Buy/Sell Protection	Description
0	0000315774	N-CSR	17	Morgan Stanley	5,000,000	Gatx Corp., 6.00%, 02/15/18	12/20/21	NaN	161,498	NaN	NaN
1	0000315774	N-CSR	17	Morgan Stanley	5,000,000	International Paper Co, 7.50%, 08/15/21	12/20/21	NaN	(44,764	NaN	NaN
2	0000883939	N-Q	09	Morgan Stanley	NaN	NaN	06/20/14	NaN	NaN	Buy	NaN
3	0001317146	N-CSRS	12	Morgan Stanley Capital Services, Inc.**	9000	NaN	6/20/16	16,340	-	Buy	NaN
4	0000837529	N-Q	07	Morgan Stanley International Limited	500000	Goldman Sachs International Hartford Financial Services Group Inc.	December 20, 2011	NaN	NaN	NaN	NaN
5	0001320615	N-CSRS	10	Morgan Stanley Capital	361,080	NR	12/20/10	NaN	NaN	NaN	NaN
6	0001320615	N-CSRS	10	Morgan Stanley Capital Services Inc	164,700	NaN	12/20/19	NaN	NaN	NaN	NaN
7	0001320615	N-CSRS	10	Morgan Stanley Capital Services Inc	$2,350,000	NaN	03/20/16	NaN	NaN	NaN	NaN
8	0001320615	N-CSRS	10	Morgan Stanley	5,000,000	NaN	12/20/15	NaN	NaN	NaN	NaN
9	0000883939	N-CSRS	10	Morgan Stanley	90,000	NaN	06/20/15	(8	(61	NaN	NaN
10	0000883939	N-CSRS	10	Morgan Stanley	NaN	NaN	12/20/15	(5	(120	NaN	NaN
11	0000315554	N-CSRS	09	Merrill Lynch International	500	Morgan Stanley /Abitibi-Consolidated, Inc.	Sep 2010	(1,980,450)	NaN	NaN	NaN
12	0000315554	N-CSRS	09	Goldman Sachs & Co.	$ 3,930	Morgan Stanley /American Airlines, Inc.	Sep 2012	(1,790,881)	NaN	NaN	NaN
13	0000315554	N-CSRS	09	Bank of America	3570	Morgan Stanley / AMR Corp.	Jun 2013	(1,943,439)	NaN	Sell	NaN
14	0000315554	N-CSRS	09	Barclays	500	Morgan Stanley /AMR Corp.	Jun 2013	(941,724)	NaN	Sell	NaN
15	0000315554	N-CSRS	09	Barclays BankAlcoa, Inc.	1500	Morgan Stanley / BoWater, Inc.			NaN	NaN	NaN
16	0001508782	N-CSR	13	JPMorgan ChaseCDX.NA.IG.20	200	Morgan Stanley /Delta Airlines, Inc.	4/24/2014	4147	4161	NaN	Put Option - OTC - Morgan Stanley Capital Services Inc., USD vs JPY
17	0001508782	N-CSR	13	JPMorgan ChaseCDX.NA.IG.20	200	Morgan Stanley /Delta Airlines, Inc.	4/24/2014	4147	4161	NaN	Put Option - OTC - Morgan Stanley Capital Services Inc., USD vs

Figure 5: Credit Default Swap Dataset with all the categorical variables

3. **CRF Classifier for CDS Reporting Dataset** - An implementation of Conditional Random Field for the CDS dataset developed as well as techniques for hyper-optimization using 3-fold cross-validation.

4. **Credit Default Swap Search Engine** - A web application to allow researchers and other users to search for credit default swap reporting by Reference Entity, Counterparty or Expiration Date.

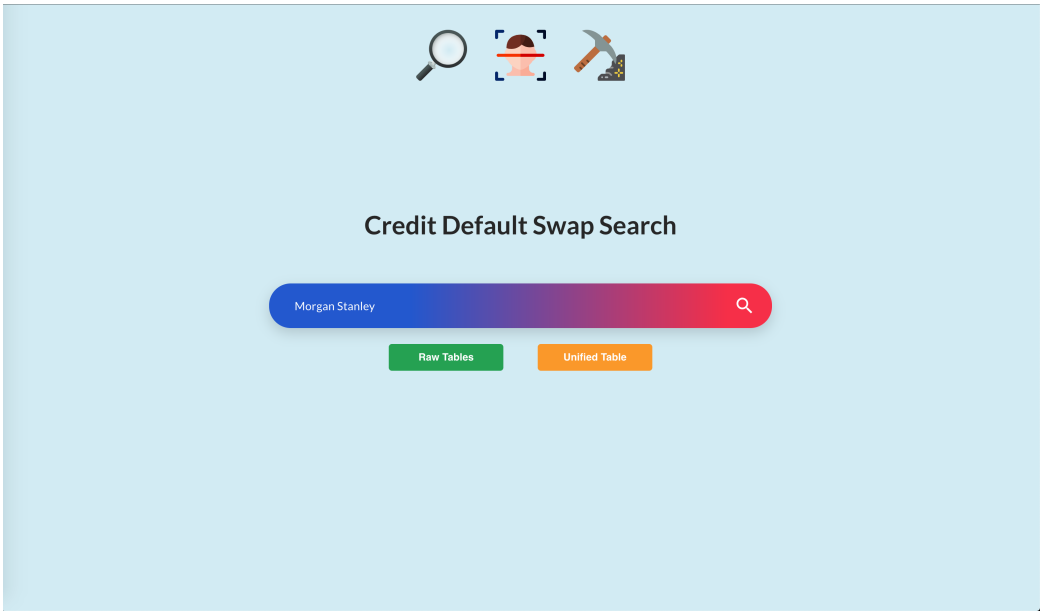

Figure 6: View of Credit Default Swap Search Engine

5. **Report Processing** - A web application built using Flask for users to upload reports containing any type of CDS information and display the Credit Default Swap that it contains.

1.5 Outline of reports

The documentation for this project has been divided into two reports. Even though both the reports share the same background and motivation, the methodology and results for each of them have been specifically written to dive deeper into the implementation and difficulties encountered for each of the two aspects. So it is imperative that the reader must go through both the reports in order to completely understand each aspect of the project.

The report going through Data Preprocessing, written by Varun Vamsi Saripalli, goes over the process of data collection, data cleaning and the tools and techniques used to achieve the Credit Default Swap reportings dataset.

This report will go over the Natural Language Processing techniques implemented to facilitate the information extraction process on unstructured sentences. First, I shall go over the literature review on the NLP techniques that we have implemented to conduct information extraction. An in-depth description will be provided for the implementation of the Conditional Random Fields on the CDS reporting dataset that we produced. Then, we analyze the results from the CRF model and provide a detailed comparison against CRF models used in finance-specific datasets to benchmark our performance. Finally, a web application featuring a Credit Default Swap Search Engine and CDS information extractor will be presented and some questions surrounding Credit Default Swaps will be answered by using the CDS dataset.

2 Literature Review

2.1 Rule Based Approach for Information Extraction

Information extraction has often been tackled with rule-based approach extraction which includes scripting with a wide range of rules accounting for every possible combination of a sentence in order to extract the required information. For example, Sheikh and Conlon, employed a rule-based approach by studying a sample space of financial documents and developing rules to extract them(Sheikh and Conlon, 2010). These efforts often led to high precision and recall immediately. However, when presented with newer formats of similar documents and dealing with high linguistic variety, the performance was lackluster. Furthermore, the amount of manual effort going into rule-based approaches can be justified for a small and known sample of documents. But they do not account of minor variations in input data which leads them to not being able to extract information when presented with new information.

However, it is important to note that it is not completely possible to eliminate rule-based approaches. Many times, this approach is required to be employed if we are dealing with structured data or data with a certain set of formats. This forms a perfect use case for Named-Entity Recognition systems, specifically sequence labeling. Since sequence labeling learns the features of the entities or words to be extracted in the sentence, it allows us to employ this technique to the kind of data where the variations in the data cannot be accounted for by rule-based approaches.

2.2 Sequence Labelling

Sequence Labelling involves the task of assigning a single label to each element in each sentence. This element could be a word or a group of words. The labels could be parts of speech tags or predefined labels such as one given in figure 7.

```
John  lives in New   York  and works for the European Union
B-PER O     O  B-LOC I-LOC O   O     O   O   B-ORG    I-ORG
```

Figure 7: Example of a sentence with its corresponding labels (Li, 2018)

Here, the entities are LOC, ORG, PER, and MISC for location, organization, person and miscellaneous. The no-entity tag is shown by O tag which means that specific element has no label. Because some entities (like New York) have multiple words, we use a tagging scheme to distinguish between the beginning (tag B-...), or the inside of an entity (tag I-...). So sequence labeling can be treated as a combination of classification tasks where the algorithm classifies each element in the sentence into a label from a predefined set of labels. The accuracy of this algorithm can be greatly improved by designing a sequence labeling algorithm which takes into account the features of its nearby elements and then classifies the current element into one of the labels. This solves our problem of rule-based methods being unable to account for variations in input data.

Sequence Labelling algorithms are mostly based on probabilistic or deep learning methods. The probabilistic method like Conditional Random Fields (CRFs) and Hidden Markov Models (HMMs) assign labels based on the probability of a particular tag sequence occurring. Deep Learning methods involve using recurrent neural networks which allow retaining contextual information of the sentence and at the same time retaining the information of distributional representations of the sentence. Due to this, deep learning

methods often prove to be more accurate in assigning labels to each element in the sentence. However, owing to the low number of the unstructured sentences in our use case, deep learning methods cannot be employed on our dataset. Therefore, probabilistic methods are a better fit for our problem statement.

2.3 Conditional Random Fields

Probabilistic classifiers mainly fall into two categories, discriminative and generative classifier models. The significant difference between discriminative and generative is that discriminative models model conditional probability distribution, i.e. P(y|X) while the generative models try to model a joint probability distribution, i.e., P(X,Y) (Sutton, 2010). Our use case requires us to account of elements nearby the current element that we trying to classify, so it is imperative for us to consider conditional probability distribution instead of joint probability distribution.

The objective of a sequence labeling problem is to find the probability of a sequence of labels(y) given an input of sequence of vectors (X). This probability is denoted by P(y|X). This makes Conditional Random Fields the perfect tool to serve our purpose.

Let's assume that the training set consists of input and target sequence pairs (X_i, y_i). The i^{th} sequence of vectors is $X_i = [x_1 ,...,x_l]$. The i^{th} target sequence of labels is $Y_i = [y_1 ,...,y_l]$ and l is the length of the sequence. For a general mathematical understanding, lets assume that for a sample (X,y), in a standard sequence labelling problem using classification we compute P(y|X) by taking the product of the probability of element at the nth position in the sequence where $1 \leq n \leq l$.

$$P(y|X) = \prod_{k=1}^{l} P(y_n|x_n) \tag{1}$$

(1) can be expanded to

$$P(y|X) = \frac{exp(\sum_{n=1}^{l} \bigcup(x_n, y_n))}{\prod_{n=1}^{l} Z(x_n)} \tag{2}$$

The expanded equation is essentially modelling $P(y_k|X_k)$ with a normalized exponential. This imitates the softmax operation widely used in neural networks and mimics its output (Lafferty & Macallum, 2001). Also here, U(x,y) is known as emissions scores which is essentially the score generated for a label y given the x vector at n^{th} timestep. The x vector in practice is the concatenation of the surrounding elements to the element that we are considering. Z(x) is known as the partition function which is normalization factor since we would want the total probability to be equal to 1 (Zhu, 2010).

So now we have established a regular sequence labelling model with a function which mimics the softmax activation to generate probabilities for each element in the input. Now, we will add further learnable weights to model the successive elements that could be present in the input. This means that we are modelling the relationship between successive labels. To implement that, we simply multiply or previous probability by $P(y_{k+1}|X_k)$ and rewrite the ***emission scores*** U(x,y) and add learnable ***transition scores*** T(x,y). This gives us

$$P(y|X) = \frac{exp(\sum_{n=1}^{l} \bigcup(x_n, y_n) + \sum_{k=1}^{l-1} T(y_k, y_{k+1}))}{\prod_{n=1}^{l} Z(x_n)} \tag{3}$$

T(x,y) is essentially a matrix where each element in it is a learnable parameter which represents the transition from the i^{th} label to j^{th} label. This gives us the

linear-chain Conditional Random Field where (3) is the conditional probability for each element(Zhu, 2010).

3 Methodology

After a comprehensive and thorough look on sequence labelling and CRF, we have made some key decisions on using linear-chain CRF to classify each element in a sentence into predefined labels. The initital process includes the data collection and cleaning which has been thoroughly described in the Data processing report by Varun Vamsi Saripalli. This section assumes that we have our training data labelled and ready to train a linear-chain CRF model.

3.1 Data Preprocessing

3.1.1 Parts of Speech Tag Generation

Parts of Speech form the building blocks of understanding the conext of each element in a sentence. For example, if we do not include POS tags in our data, our trained model will not be able to capture the difference between "I like a potato" and "I am like Helen" where the former has a verb context while the latter has a preposition context. Furthermore, POS tags have been deemed to be useful in extracting relations between words and also building lemmatizers to reduce a word its root form. However, in our case, we are using POS tags to extract relationship between consecutive words.

In order to do so, we are using the NLTK library provided as an open source library Stanford's CoreNLP API. In order to tag our words, we simply call the *pos_tag(word)* to generate a POS tag the word in our tagged dataset. Once, we do that for the entire corpus of sentences, our dataset would look as shown in Figure 8.

	Token	NE	POS
0	50,00,000	B-Notional Amount	CD
1	USD	O	NNP
2	10/15/03	B-Expiration Date	CD
3	Agreement	O	NNP
4	with	O	IN
5	Deutsche	B-Counterparty	NNP
6	Bank	I-Counterparty	NNP
7	AG	I-Counterparty	NNP
8	dated	O	VBD
9	1/21/03	O	CD

Figure 8: Generated POS tags with other columns in the data

3.1.2 Filtering Labels

As shown in figure 8, the words have a label associated to them which was generated during the data preprocessing step using the Text Annotation Tool. These labels serve as one of the inputs into the Conditional Random Field model that we will implement to develop classifier to label words.

Label	**Count**
B-Counterparty	491
B-Direction of Trade	504
B-Expiration Date	492
B-Fixed Rate	511
B-Notional Amount	488
B-Reference Entity	498
I-Counterparty	843
I-Expiration Date	97
I-Fixed Rate	1
I-Notional Amount	2
I-Reference Entity	1100
O	15094

Table 1: Label Distribution in our data

However, from table 1 we can infer that almost 15,094 labels are of the O label or no-entity label which means these words have no significance in our analysis as they do not carry information that we want to extract. This could bring in class imbalance and would inflate the accuracy of our CRF model as it would tend to classify other label as O(no-entity label). This would lead to a high theoretical accuracy but it would classify most words as O label.

To mitigate this, we could either oversample the other non-O labels or undersample the O-label. We chose to undersample the O-label by simply dropping the words which carry the label O. So the final labels that we will classify all the words into are : *B-Direction of Trade, B-Reference Entity, I-Reference Entity,B-Fixed Rate, B-Counterparty, I-Counterparty, B-Expiration Date, I-Expiration Date, B-Notional Amount, I-Notional Amount and I-Fixed Rate.*

3.2 Feature Extraction

In section 2.1.1, we generated POS tags for every word in our corpus of sentences. This is because they convey important information about the word or group of words in a sentence like its semantic meaning and its position in the sentence. This would allow us to develop feature function which is one of the significant aspects of CRF. As we are essentially building linear-chain CRF, the feature function would look like:

$$f_i(z_{n-1}, z_n, x_{1:N}, n) \tag{4}$$

where z_{n-1}, z_n are adjacent states or words in a sentence and the whole sentence sequence is denoted by $x_{1:N}$. Let's take an example specific to our use case to further understand it's significance.

Let's assume that a simple feature function which produces binary values for if the current word is *JPMorgan* and the current label is *B-Counterparty*. The CRF model will utilize this feature function with it's corresponding weight λ_1 . In this case, if $\lambda_1 > 0$ and if the current word is *JPMorgan* and current label is *B-Counterparty* then our feature function will be active. This is interpreted by the CRF model as increase in probability of labelling a word as *B-Counterparty* if the word is *JPMorgan*. Similarly, if $\lambda_1 < 0$, then the CRF model will have a lower probability of tagging a word as *B-Counterparty* if the word is *JPMorgan*. Now the λ_1 can either be specified through the process of labelling or learning from the corpus or both. In our use case, we will learn $\lambda_{1:n}$ from training data, for which we developed feature dictionaries from our traning data for the CRF model to train on.

Since, we are not able to implement LSTMs to extract the word features and pass it to successive units ahead, there is a need for feature dictionary which would consist of word features of every element that would be extracted from

each sentence. A list of feature dictionaries for each word token in a sentence can be extracted, corresponding to a list of labels for each word token in a sentence.

1. Previous Parts of Speech Tag
2. Current Parts of Speech Tag
3. Word.isdigit() [True if word is a number false otherwise]
4. Word.isUpper() [True if word is a upper false otherwise]
5. Previous Word
6. Word.isLower() [True if word is a lower false otherwise]
7. Word.isLower() [True if word is in title case false otherwise]

```
First token features:
---------------------
{'Token.lower()': 'upon', 'Token.isupper()': False, '+1:Token.lower()':
'a', 'Token[-2:]': 'on', '+1:Token.istitle()': False, 'Token[-3:]': 'po
n', 'POS[:2]': 'IN', 'POS': 'IN', 'Token.istitle()': True, '+1:POS[:2]':
'DT', 'BOS': True, '+1:Token.isupper()': False, '+1:POS': 'DT', 'bias':
1.0, 'Token.isdigit()': False}
```

Figure 9: Example of Feature Dictionary for the word 'upon'

Let's visualize a feature dictionary for the word *upon* in figure 9. Using a feature dictionary like this for each word, we synthesized a list of feature dictionaries for all the words in the data. Furthermore, we have a used standardized format for the feature dictionary as dictated in *python-crfsuite* documentation in order to ensure easy replication of similar results.

3.3 Training the Conditional Random Field Model

Before,we could start training our Conditional Random Field model, we had to choose the right implementation of linear-chain CRF that would be used. We selected a fast C++ implementation of CRF called *CRFSuite* which has proven state-of-the-art results for sequence labelling and named entity recognition problems. It's other features included:

1. State-of-the-art training method using methods like Limited-methods BFGS
2. Linear-chain (first-order Markov) CRF
3. Performance evaluation on training

Following this, using predefined functions, X and y were denoted as a list of feature dictionaries for each word in each sentence and as a list of labels for each word in each sentence. Making use of *scikit-learn* library's *test_train_split* function, the data was split into training and testing data with 80% of the data dedicated for training while 20% for testing.

Serving as an extension to the content covered in the literature review for CRF, the partition state or normalization factor used to compute the probability in the end which is denoted by Z in equation(3) can be expressed as following:

$$Z(X) = \sum_{y_1^{'}} \sum_{y_2^{'}} ... \sum_{y_n^{'}} exp(\sum_{k=1}^{l} U(x_k, y_k^{'}) + \sum_{k=1}^{l-1} T(y_k^{'}, y_{k+1}^{'}))) \tag{5}$$

However, the computation of the partition function Z is computationally intensive as it has a lot nested loops (Zhu, 2010) . There are $l!$ computations required over the label set. This gives us a total complexity of $O(l!\,|y|^2)$. However, *CRFSuite* library makes it easy to tackle this by providing the use of *forward or backward algorithm* as simple function argument while

training a CRF model. Depending on the order of iteration for a sequence, we can choose the algorithm we want to use. Given that we are employing Linear-Chain CRF, we would use forward-backward algorithm. Finally, for optimization, standard optimization algorithms like Stochastic Gradient Descent or Limited-Memory BFGS could be used. We used L-BFGS as the library *CRFSuite* provided support for the it.

So in order to express our problem in a mathematical form, let us assume that our fully labelled data is represented as $(w^{(1)}, t^{(1)}, s^{(1)})..., (w^{(n)}, t^{(n)}, s^{(1)})$, where $w^{(i)} = w^{(i)}_{1:N}$ are the sequence of words present in our unstructured sentences containing Credit Default Swaps, $t^{(i)} = t^{(i)}_{1:N}$ are labels for each corresponding word in the sentence, and $s^{(i)} = s^{(i)}_{1:N}$ are the corresponding parts-of-speech tag for the corresponding word, respectively.

We already know that in CRFs, the objective of parameter learning is to maximize the conditional likelihood on the basis of training data. This can be represented as:

$$\sum_{j=1}^{M} logp(t^{(j)}|w^{(j)}, s^{(j)}) \tag{6}$$

In order to stop over-fitting, we conduct penalization on log-likelihood with a zero-mean Gaussian Distribution over the parameters. This makes equation (6) as

$$\sum_{j=1}^{M} logp(t^{(j)}|w^{(j)}, s^{(j)}) - \sum_{i}^{F} \lambda_i^2/2\sigma^2 \tag{7}$$

As equation (7) is concave in nature, we can deduce that λ would have unique set of optimal values. With the help of L-BFGS's gradient, we learn the parameters. So training the CRF model would allow us to find the optimal values of λ for the training data.

3.4 Credit Default Swap Search Engine

Upon extracting information from both structured and unstructured formats of Credit Default Swap reportings, we developed a search engine to enable future research studies to further take advantage of the consolidated data that has been aggregated through rule-based as well as NLP techniques.

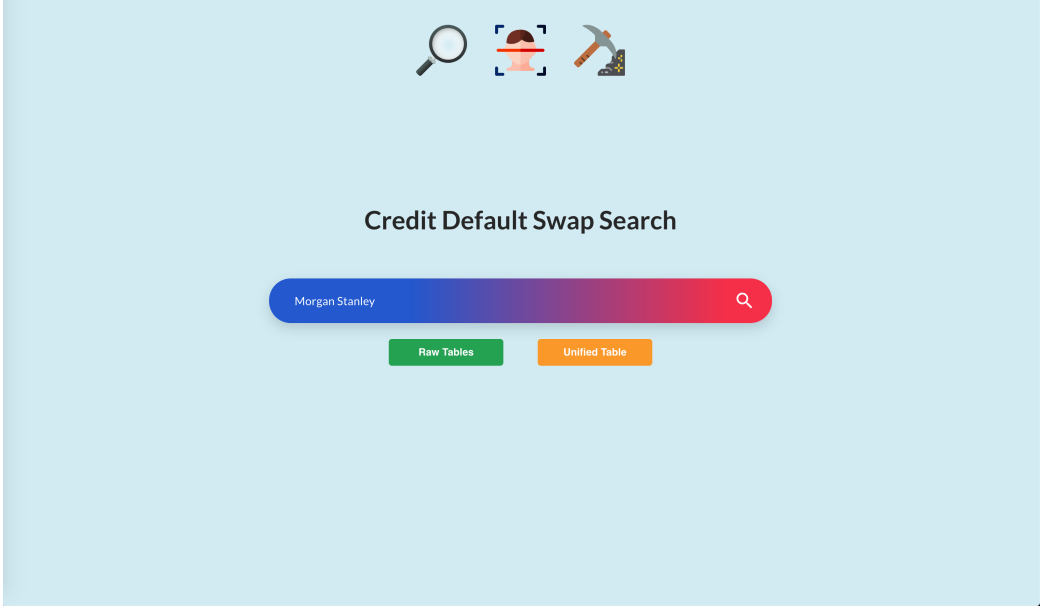

Figure 10: Final Credit Default Swap Search Engine

This web application was built on Flask with the entire dataset of 16,813 rows into an array of JSON objects. JSON objects are the defacto standard for query based searching and also allow swift query and return time. Furthermore, this web application also serves as a way for researchers and financial analysts to upload reports that they want the Credit Default Swap information extracted from. The application itself is capable of extracting both structured as well as unstructured reporting of CDS as the model is running at the backend and is served as a RESTful framework.

4 Results and Analysis

4.1 CRF Results

4.1.1 CONLL2003 Dataset Performance

In order to ensure that the readers get a metric of what kind of performance our CRF model is capable of, we first trained it on CONLL 2003 dataset. The CONLL 2003 dataset is a standard for benchmarking models in sequence labelling tasks and would allow us to easily compare with other similarly implemented model for the same training data.

We provide numerical results in the form of F1-Score, Precision and Recall. Precision here is defined as the fraction of relevant data points from the retrieved data points while Recall is defined as the fraction of relevant data points returned from the total number of relevant data data points. So precision and recall will serve as a good metric to understand and measure the relevance of our results ("Precision-Recall scikit-learn 0.20.3 documentation," n.d.).

Therefore, F1 score is the measure of a test's accuracy and is defined as the weighted harmonic mean of the precision and recall of the test. With this, we first trained the CRF Model on the CONLL 2003 dataset and tried to validate performance by benchmarking with other state of the art implementations. Benchmarking allows us to validate the conditions and settings that we have used to train our CRF model. From Table 2, we can incur that our CRF model closely represents the Conv-CRF(Collobert et al., 2011) in terms of both the settings used for traning as well as accuracy.

System	**Accuracy**
Conv-CRF (Senna + Gazetteer)(Collobert et al., 2011)	89.59%
Early CRF Models (MacCullum, Li (2005))	84.04%
Conv-CRF(Collobert et al., 2011)	81.47%
CRF with Lexicon Infused Embeddings (Passos et al., 2014)	**90.90%**
CRF (Our)	**81.21%**

Table 2: Benchmarking performance with other state-of-the-art CRF models

4.1.2 Performance on Unstructured CDS Reporting

As mentioned in the methodology, the next step was to train our CRF model for the unstructured Credit Default Swap reporting from 2004-2017. This primarily includes sentences like those shown in figure (3). The training itself took about 20 minutes on a 2.3GHz i7 processor and returned with following results shown in table 3.

Label	**Precision**	**Recall**	**F1-Score**	**Support**
B-Notional Amount	0.98	0.94	0.96	99
B-Expiration Date	0.96	0.97	0.97	102
B-Counterparty	0.98	0.98	0.98	101
I-Counterparty	0.97	0.96	0.96	182
B-Direction of Trade	0.96	0.97	0.97	106
B-Fixed Rate	0.98	1.00	0.99	105
B-Reference Entity	0.98	0.97	0.98	104
I-Reference Entity	0.96	0.93	0.94	245
I-Expiration Date	0.94	0.89	0.92	19
I-Notional Amount	0.00	0.00	0.00	1
I-Fixed Rate	0.00	0.00	0.00	1
Micro Average	**0.97**	**0.96**	**0.96**	**1065**
Macro Average	**0.79**	**0.78**	**0.79**	**1065**
Weighted Average	**0.97**	**0.96**	**0.96**	**1065**

Table 3: Precision, Recall and F1-Score of the CRF model on unstructured CDS reporting with scores for each label

One interesting aspect to note is that there were certain labels like *I-Notional Amount* and *I-Fixed Rate* which had only one instance in the entire dataset. Therefore, our model rejects the labels with only one instance and reports an F1-score of 0.

This allows us to establish that the CRF model that we have developed and trained on the unstructured CDS report could be used to extract key information which is often hidden in the context of unstructured sentences. However, it could be concluded that there is no prior work conducted on extracting unstructured reportings of CDS and therefore making it difficult to benchmark our performance with other studies

Studies Conducted	**F1 Score**
(Alvarado, Verspoor and Baldwin, 2013)	0.827
(Wang, Xu, Liu, Gui, and Zhou, 2015)	0.857
Bankruptcy Prediction using CRF	0.859
Our Implementation	**0.96**

Table 4: Studies published in the finance domain which used CRF to extract information and their reported F1 Scores

But we chose benchmark our performance with similar studies which implemented CRF on finance-specific datasets. we have gone over three major studies in this direction mentioned in table 4. These studies go through studies conducted on similar financial documents such as loan agreements and contracts using Conditional Random Field to extract unstructured information.

4.2 Optimizing Hyperparameters

Under this section, we experimented with different values of C1 and C2 values for the elastic net regularization. In order to achieve this we used cross-validated randomized search. To avoid a computationally intensive task, we limited the iterations to 50 and use a 3-fold cross-validation. This in turn would mean that we essentially trained a 150 models.

Following the optimization, we noticed that lower values (increased regularization strength) for both C1 and C2 values result in the best performing model - particularly for C1. After optimizing the hyperparameters, the CRF model was be evaluated again. The results of the new model have been disclosed in table 5.

Best Score	**C1**	**C2**
96.81%	0.001	0.001

Table 5: Best Score and Best parameters after conducting Randomized-CVSearch

We noticed that there isn't a significant improvement in the F1-score of the trained model after optimization. However, this is attributed to the size of data that we began with. We firmly believe that if the training data was much larger then there would have been a considerable affect of the hyperparamter optimization.

4.3 Analysis

One of the main goals of this project was to answer some questions surrounding Credit Default Swap reportings and the data processing and extraction process was supposed to allow us to have enough data to answer questions like

1. What were the trends of Credit Default Swap reporting before financial crisis, during and afer the financial crisis ?
2. What are the patterns in usage of index CDS, Sovereign CDS and Single-name CDS ?
3. Do funds who have a more structured format of reporting CDS have a higher profit margin as compared to funds who used unstructured sentences to report CDS ?

In order, to answer these questions, the Credit Defautl Swap dataset which was one of the deliverables of this project is supposed to visualize the trends in CDS reporting in a graphical manner.

The first question deals with the reporting trends of CDS from the time period of 2004-2017 and after performing a simple groupby by column name "Reporting Year" and the trends have been reported in figure 10.

We notice that the CDS reporting from the time period of 2004-2008 have an exponential rising trend. This trend could be attributed to the housing market bubble in the United States. Housing market was believed to be one of the safest and risk-free investments from a consumer standpoint. Using this trend in the market, a lot of financial institutions bought Credit Default Swap in the expectation of earning more profits through them as the thought of the financial markets collapsing was simply considered to be least probable.

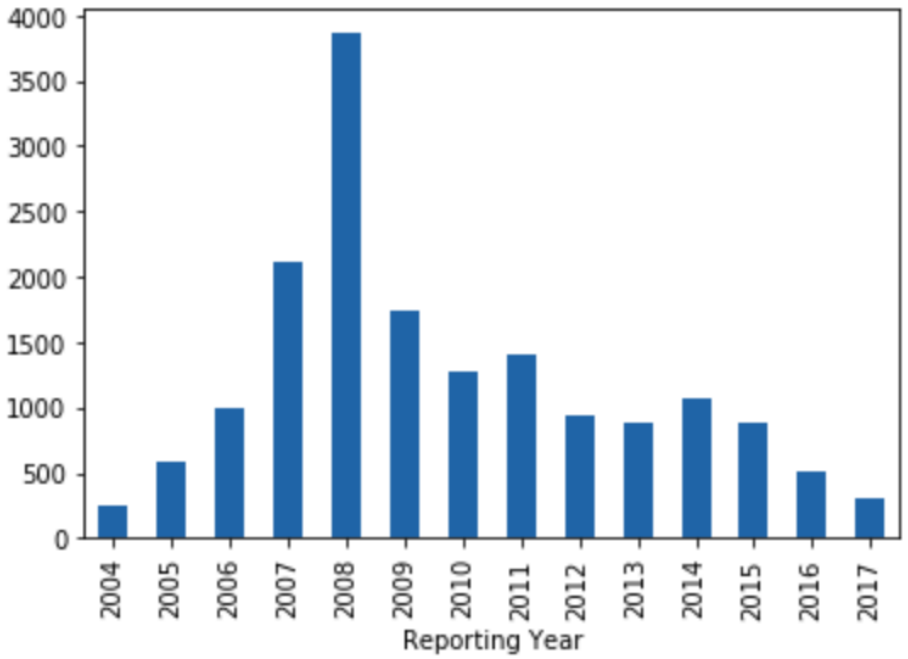

Figure 11: Trend of CDS reporting for every year from 2004-2017

However, as 2008 came along and when the housing market collapsed we see significant drop in the CDS reporting and from there on it has been going down to a record low in 2017. This mirrors the real-world circumstances as with heavy regulation from the SEC and the housing market not deemed as safe as it was at its peak, the financial institutions investing in CDS has slowly gone down and they have now started looking towards alternate credit derivatives which allows them to hedge their risk and diverse their protfolios in a better method.

The second question dives deeper into the trend of different types of Credit Default Swaps being used during 2004-2017 time period. We notice all three types of CDS were again very popular in the 2004-2008 period when the market for Credit Default Swaps was booming.

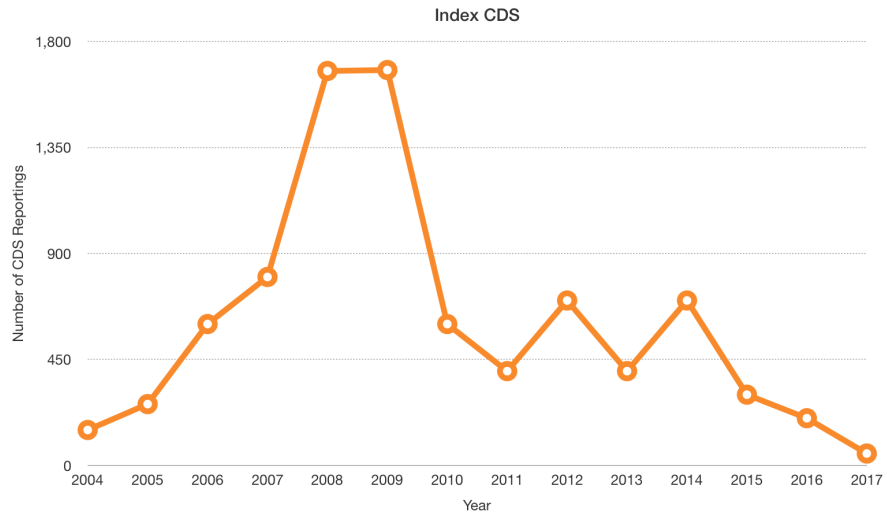

Figure 12: Trend of CDS reporting by Index CDS

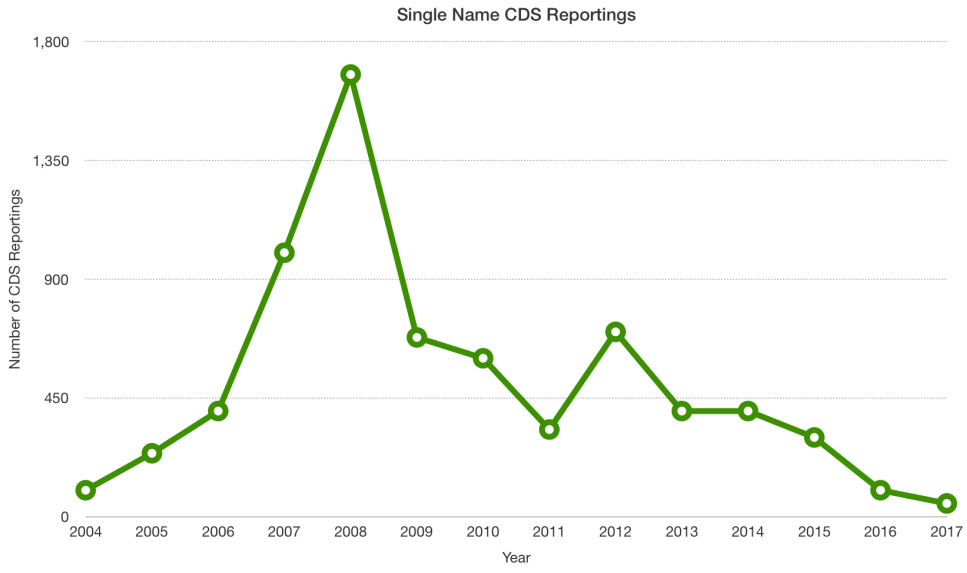

Figure 13: Trend of CDS reporting by Single Name CDS

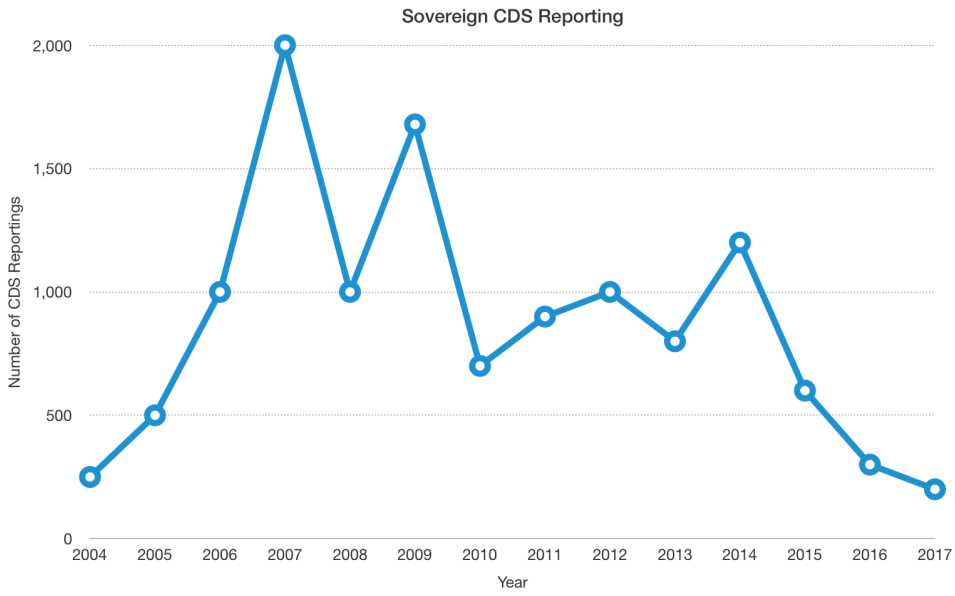

Figure 14: Trend of CDS reporting by Sovereign CDS

For sovereign CDS, we noticed a similar trend in reporting with Reference Entities including countries like China and Russia were one of the most reported Credit Default Swaps. Furthermore, Single Name and Index Name CDS, both represent a similar reporting pattern throughout the years as it clearly correlates with the rise and the fall of the CDS market.

The third question involves the comparison of structured and unstructured reporting to find if funds practicing unstructured method of reporting earn higher margin on CDS or the funds with the structured method reporting. Answering this question would require significant analysis on both structured and unstructured information, however, as explained in the section earlier, we only 1,200 reporting of unstructured CDS information from the total 16,813.

With this size of unstructured data, it is not possible to conclude that the funds

with unstructured method reporting have a higher margin of profit as the sample space for comparison is too small. However, with more unstructured CDS reporting, this could be made possible.

5 Conclusion

To summarize, we explored the importance having a consolidated database of Credit Default Swap reporting and how it can be useful for financial institutions to base their investment decisions of if they want to enter the CDS market. Moreover, this data is required for to conduct further analysis on the 2008 financial crisis and could help researchers and analysts derive new insights which could then power new conclusions. However, we also understood that we require both rule-based extraction techniques as well as natural language processing techniques to extract CDS reporitng which have been reported in different formats.

With this intention, we first proposed a framework that would use the current state-of-the-art technology to extract information from unstructured CDS reporting. After doing a thorough extraction of raw data, we understood that we do not have enough unstructured reporting to take a deep-learning approach to train a model. So we switched to training a Conditional Random Field Classifier and developed a feature function to extract the word-features of preceding and succeeding units in a sentence. Once we trained our CRF model with the tuned hyperparameters, we benchmarked our results against similarly studies conducted and compared our performance to give the reader an overview of the current state of implementation in the area of finance-specific Named Entity Recognition.

Then we also established that the NLP processed sentences will be added to the dataset of Credit Default Swap reporting from 2004-2017 which we used to answer some questions surrounding Credit Default Swap. Our analysis powered by the data we had extracted showed us that the insights we drew were in line with the real word understanding of the Credit Default Swaps on the US Economy and the 2008 Financial Crisis. The trends clearly reflected an exponential rise in CDS reporting from 2004-2008 and a step decline

thereafter. However, due to lack of unstructured method of CDS reporting, we could not analyze the effects of structured or unstructured reporting on the margins of financial institutions trading Credit Default Swaps.

Finally, as one of the significant outcomes of this project, we developed a Credit Default Swap Search Engine which is supposed to allow researchers and analysts interested in learning about a specific Credit Default Swap or a financial institution's dealings, could simply search for it through our search engine. Moreover, the Credit Default Swap dataset is supposed to enable financial analysts and researchers to further draw insights from the Credit Default Swap reporting. Finally, we developed a data extraction tool for users interested in extracting both structured and unstructured information in a financial report by simply uploading it on our web application and being able to view the results instantly.

References

A.Guettler ,T.Adam. Pitfalls and Perils of Financial Innovation: The Use of CDS by Corporate Bond Funds. Jan 10, 2015.

A. Passos, V. Kumar, and A. McCal-lum. 2014.Lexicon Infused Phrase Embeddings forNamed Entity Resolution. Proceedings of CoNLL.

Bill Y.Lin, Frnk F.Xu, Zhiyi Luo and Kenny Q.Zhu, Multi-channel *"BiLSTM-CRF Model for Emerging Named Entity Recognition in Social Media*, Shanghai, Sept 2017

Credit Default Swaps (n.d). PIMCO Funds. Available at: `https://global.pimco.com/en-gbl/resources/education/understanding-credit-default-swaps`

Li, S. (2018, August 27). Named Entity Recognition and Classification with Scikit-Learn. Retrieved April 14, 2019, from `https://towardsdatascience.com/named-entity-recognition-and-classification-with-scikit-learn-f05372f07ba2`

Precision-Recall scikit-learn 0.20.3 documentation. (n.d.). Retrieved April 14, 2019, from `https://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html`

P. Wayne (2018, October 6). Credit Default Swaps: An Introduction [Online]. Available: https://www.investopedia.com/articles/optioninvestor/08/cds.asp

R. Collobert, J. Weston, L. Bot-tou, M. Karlen, K. Kavukcuoglu and P. Kuksa.2011.Natural Language Processing (Almost) fromScratch.Journal of Machine Learning Research(JMLR)

Stanford CoreNLP Natural language software. (n.d.). Retrieved from https://stanfordnlp.github.io/CoreNLP/

Sutton, C. (2012). An Introduction to Conditional Random Fields. Foundations and Trends in Machine Learning, 4(4), 267-373. doi:10.1561/2200000013

Susmitha Wunnava, Xiao Qin, Tabassum Kakar, Elke A. Rundensteiner and Xiangnan Kong, *Bidirectional LSTM-CRF for Adverse Drug Event Tagging in Electronic Health Records*, Worcester, 2018

W.Jiang, Z.Zhu. Mutual Funds Holdings of Credit Default Swaps. September 2016.

Wang, B. (2018). Forecasting Bankruptcy Prediction using Conditional Random Fields (Unpublished doctoral dissertation). Ghent University.

Wang, S., Xu, R., Liu, B., Gui, L., Zhou, Y. (2014). Financial named entity recognition based on conditional random fields and information entropy. 2014 International Conference on Machine Learning and Cybernetics. doi:10.1109/icmlc.2014.7009718

Zhiheng Huang, Wei Xu, and Kai Yu, *Bidirectional LSTM-CRF Models for Sequence Tagging*, August 2015

Zhu, X. (2010). CS769 Spring 2010 Advanced Natural Language Processing. Retrieved from http://pages.cs.wisc.edu/~jerryzhu/cs769/CRF.pdf