Understanding Financial Reports using NLP

Final Presentation

Supervisor: Dr. Ruibang Luo

Tarun Sudhams
Varun Vamsi Saripalli

Outline

- Motivation
- Deliverables
- Methodology
- Analysis
- Predicting Financial Health of a Company
- Conclusion

Motivation

- CDS considered to be one of the culprits of the 2007-2008 Financial Crisis
- Huge volumes of unstructured CDS data present
- Lack of public information regarding CDS trading
- Extracting and analyzing such information is thus extremely important
- Making the data available and open source will help future research and companies looking to use CDS

Deliverables

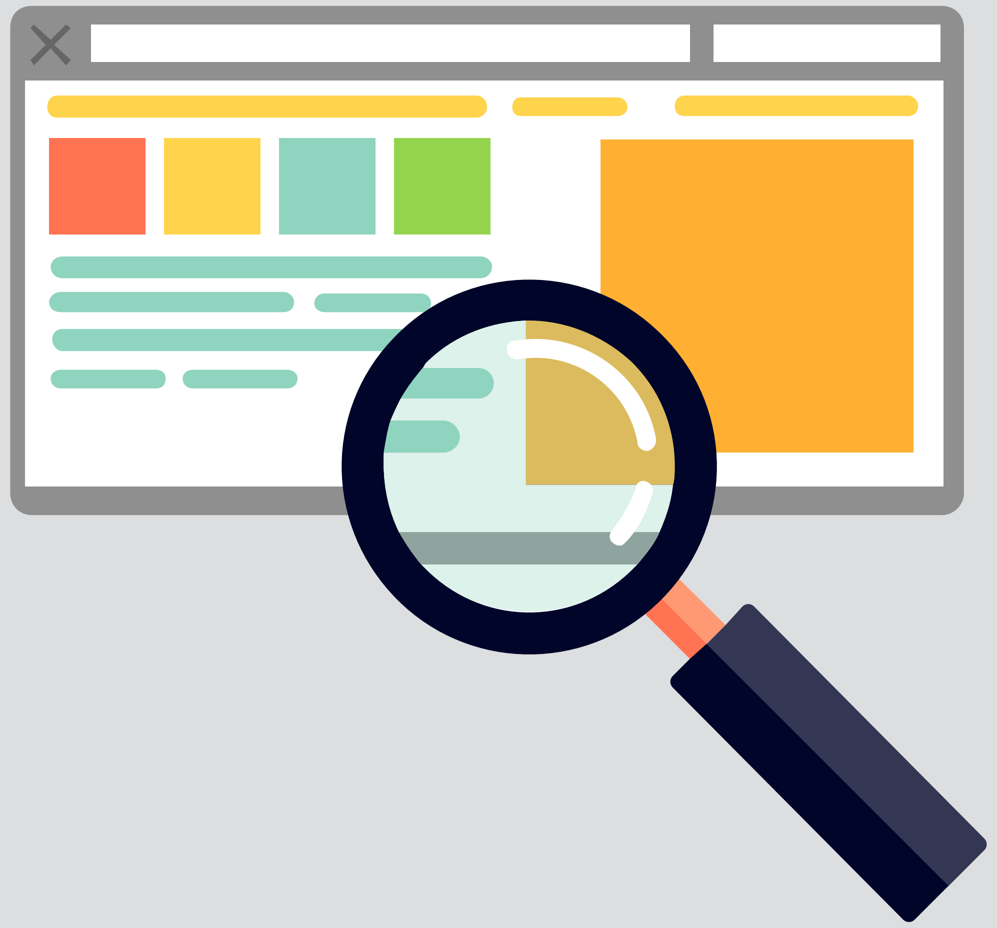
- Text Annotation Tool
- Credit Default Swaps Reportings Dataset
- CRF Tagger for CDS Reportings Dataset
- Credit Default Swap Search Engine
- Report Processing

Methodology

Approaches

- Rule Based Extraction of Data
  - Created a python script to extract tables from the reports.
  - Accuracy of the script is 91.33%
  - Merged the Data to create a database containing all CDS information

- Natural Language Processing
  - Discovered that unstructured data was scarce.
  - Deep Learning not possible with the amount of data we had.
  - Decided on a CRF Model for NLP
  - Tagged data for CRF training



Rule-Based Extraction Method

Step1: Data Collection and Cleaning

Scraped SEC reports from the United States SEC website → Used python with web scraping frameworks to get the data → Crawled 146GB of financial reports dating from 2004-2017

Step1: Data Collection and Cleaning

Scraped SEC reports from the United States SEC website

Used python with web scraping frameworks to get the data

Crawled 146GB of financial reports dating from 2004-2017

Step2: Extraction and Formatting tables

Python Script run on reports to extract CDS tables

Step1: Data Collection and Cleaning

Scraped SEC reports from the United States SEC website

Used python with web scraping frameworks to get the data

Crawled 146GB of financial reports dating from 2004-2017

Step2: Extraction and Formatting tables

Python Script run on reports to extract CDS tables

Removed tables containing non-CDS information by checking for keywords

Step1: Data Collection and Cleaning

Scraped SEC reports from the United States SEC website

→

Used python with web scraping frameworks to get the data

→

Crawled 146GB of financial reports dating from 2004-2017

Step2: Extraction and Formatting tables

Python Script run on reports to extract CDS tables

→

Removed tables containing non-CDS information by checking for keywords

→

Format the tables to remove unwanted characters and align columns

Step2.1 Script to extract tables

Information concerning the credit default swap agreements outstanding for the ING VP Balanced Portfolio at June 30, 2006 is shown below:

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	$	263,000	$	(10,849)
UBS AG	CDX.NA.IG.6	Buy	(0.40)%	6/20/11	$	10,157,000		(13,675)
Morgan Stanley	CDX.NA.HY.6	Buy	(3.45)%	6/20/11	$	1,025,000		2,222
Morgan Stanley	CDX.NA.HY.6	Buy	(3.45)%	6/20/11	$	1,025,000		(4,800)
UBS AG	CDX.NA.HY.6	Buy	(3.45)%	6/20/11	$	1,025,000		(18,083)
Morgan Stanley	CDX.NA.HY.6	Buy	(3.45)%	6/20/11	$	1,025,000		(4,061)
Citibank N.A.	Windstream 8.125% due 8/1/2013	Buy	(1.60)%	9/20/11	$	446,000		(2,449)
UBS AG	Windstream 8.125% due 8/1/2013	Buy	(1.63)%	9/20/11	$	512,500		(3,083)
							$	(54,778)

CDS in a formatted report

Step2.1 Script to extract tables

table2.csv - LibreOffice Calc (on bal2)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Counterparty	�	Reference Entity/Obligation	�	Buy/Sell Protection	�	(Pay)/Receive Fixed�Rate	�	Expiration Date	�	Notional Amount	�	Unrealized Appreciation/ (Depreciation)	�		
2	Citibank N.A.	�	Georgia Pacific 8.125% due 5/15/2011	�	Buy	�	(3.55	)%	�	12/20/10	�	$	263000	�	$	(10,849
3	UBS AG	�	CDX.NA.IG.6	�	Buy	�	(0.40	)%	�	6/20/11	�	$	10157000	�	(13,675	)
4	Morgan Stanley	�	CDX.NA.HY.6	�	Buy	�	(3.45	)%	�	6/20/11	�	$	1025000	�	2222	�
5	Morgan Stanley	�	CDX.NA.HY.6	�	Buy	�	(3.45	)%	�	6/20/11	�	$	1025000	�	(4,800	)
6	UBS AG	�	CDX.NA.HY.6	�	Buy	�	(3.45	)%	�	6/20/11	�	$	1025000	�	(18,083	)
7	Morgan Stanley	�	CDX.NA.HY.6	�	Buy	�	(3.45	)%	�	6/20/11	�	$	1025000	�	(4,061	)
8	Citibank N.A.	�	Windstream 8.125% due 8/1/2013	�	Buy	�	(1.60	)%	�	9/20/11	�	$	446000	�	(2,449	)
9	UBS AG	�	Windstream 8.125% due 8/1/2013	�	Buy	�	(1.63	)%	�	9/20/11	�	$	512500	�	(3,083	)
10	�	�	�	�	�	�	�	�	�	�	�	�	$	(54,778	)	
11																

Extracted Table in Raw Form

Step2.2 Removing non-CDS tables

table3.csv - LibreOffice Calc (on bal2)

	A	B	C	D	E	F	G	H	I	J	K	L
1												
2	Fund	��	Maximum��Amount	��	��	Maximum��Amount	��					
3												
4	Global Alternatives Fund	��				��						
5	Over-the-counter counterparty credit risk	��				��						
6	Forward foreign currency contracts	��	$	14098719	����	��	$	7163923	����			
7	Collateral pledged to UBS AG	��	��	29530000	����	��	��	29530000	����			
8		��	��	��	��	��	��	��	��			
9	Total over-the-counter counterparty credit risk	��	��	43628719	����	��	��	36693923	����			
10		��	��	��	��	��	��	��	��			
11	Exchange traded counterparty credit risk	��				��						
12	Futures contracts	��	��	39564404	����	��	��	39564404	����			
13	Margin with brokers	��	��	119979543	����	��	��	119979543	����			
14		��	��	��	��	��	��	��	��			
15	Total exchange traded counterparty credit risk	��	��	159543947	����	��	��	159543947	����			
16		��	��	��	��	��	��	��	��			
17	Total counterparty credit risk	��	$	203172666	����	��	$	196237870	����			
18		��	��	��	��	��	��	��	��			

Example of Non-CDS table

Step2.2 Removing non-CDS tables

table4.csv - LibreOffice Calc (on bal2)

	A	B	C	D	E	F	G	H	I
1	Pay/ReceiveFloating Rate	Floating Rate Index	FixedRate	MaturityDate	NotionalAmount	MarketValue	UnrealizedAppreciation/(Depreciation)	Variation Margin	
2	Receive	3-Month USD-LIBOR	2	06/18/2019	165800	(5,097	(3,900	132	0
3	Pay	3-Month USD-LIBOR	2.25	12/17/2019	89600	3692	1383	0	(76
4	Receive	3-Month USD-LIBOR	3.75	09/17/2043	209000	(51,699	(36,613	734	0
5	Pay	3-Month USD-LIBOR	3.5	06/19/2044	199700	44515	51030	0	(700
6	Receive	3-Month USD-LIBOR	3.25	06/17/2045	22800	(3,541	(1,286	79	0
7	Pay	6-Month AUD-BBR-BBSW	3.5	06/17/2025	7600	265	77	0	(44
8									
9									

Example of Interest-Rate Swaps

Step2.3 Formatting extracted table

	A	B	C	D	E	F	G	H
1	Counterparty	Reference Entity/Obligation	Buy/Sell Protection	(Pay)/Receive Fixed Rate	Expiration Date	Notional Amount	Unrealized Appreciation/ (Depreciation)	
2	Citibank N.A.	Georgia Pacific 8.125% due 5/15/2011	Buy	(3.55	12/20/10	263000	(10,849	
3	UBS AG	CDX.NA.IG.6	Buy	(0.40	6/20/11	10157000	(13,675	
4	Morgan Stanley	CDX.NA.HY.6	Buy	(3.45	6/20/11	1025000	2222	
5	Morgan Stanley	CDX.NA.HY.6	Buy	(3.45	6/20/11	1025000	(4,800	
6	UBS AG	CDX.NA.HY.6	Buy	(3.45	6/20/11	1025000	(18,083	
7	Morgan Stanley	CDX.NA.HY.6	Buy	(3.45	6/20/11	1025000	(4,061	
8	Citibank N.A.	Windstream 8.125% due 8/1/2013	Buy	(1.60	9/20/11	446000	(2,449	
9	UBS AG	Windstream 8.125% due 8/1/2013	Buy	(1.63	9/20/11	512500	(3,083	
10								
11								

Formatted CDS Table

Credit Default Swap Dataset

Rows - 16,813

	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

Performance: Table Extraction

Type of Report	Total Number of Files in Corpus	Total Extracted as Structured
N-CSR	7750	5198
N-CSRS	3304	2168
N-Q	3670	2720

Number of reports with structured data (tables)

Type of Report	Total Extracted as Structured	No of Reports with useful CDS Data
N-CSR	5198	2016
N-CSRS	2168	724
N-Q	2720	1365

Number of reports with structured data (tables) with CDS

Type of Report	Number of Reports Checked	Data Accurately Extracted
N-CSR	100	93
N-CSRS	100	89
N-Q	100	92

Accuracy: 91.33%

Observations

Agreement with Bear Sterns and Co., dated 11/2/05 to receive monthly the notional amount multiplied by 2.10% and pay in the event of a write down, failure to pay a principal payment or an interest shortfall on BSCMS 2005-PWR9 K.

Unstructured Methods of reporting Credit Default Swap

- Need to prepare unstructured data for Natural Language Processing
- Run a python script to extract out only the useful unstructured sentences from the files
- Manually tag the data using our Text Annotation Tool
- Text Annotation Tool automatically formats tagged dataset into desired format

Natural Language Processing

Steps for conducting NLP

Parts of Speech Generation

- Parts of Speech form the building blocks of understanding the context of each element in a sentence.

- POS tags have been deemed to be useful in extracting relations between words and also building lemmatises to reduce a word to its root form.

- NLTK Library provided by Stanford's CoreNLP API used to tag parts of speech.

	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
10	to	O	TO
11	pay	B-Direction of Trade	VB
12	11.20%	B-Fixed Rate	CD

Filtering Labels

- 15,094 labels O-label or no-entity label. No significance in our analysis as they do not carry information that we want to extract.

- Could lead to class imbalance and not give accurate results. This would lead to a high theoretical accuracy but it would classify most words as O label.

- We chose to undersample the O-label by simply dropping the words which carry the label O

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

Feature Extraction and Training

- Feature function is a function that takes in as input:a sentence, the position of a word in the sentence, the label of the current word, the label of the previous word

$$score(l|s) = \sum_{j=1}^{m} \sum_{i=1}^{n} \lambda_j f_j(s, i, l_i, l_{i-1})$$

- Then we assign each feature function a weight and simply add up all the weights for each word to generate a score

$$p(l|s) = \frac{exp[score(l|s)]}{\sum_{l'} exp[score(l'|s)]} = \frac{exp[\sum_{j=1}^{m} \sum_{i=1}^{n} \lambda_j f_j(s,i,l_i,l_{i-1})]}{\sum_{l'} exp[\sum_{j=1}^{m} \sum_{i=1}^{n} \lambda_j f_j(s,i,l'_i,l'_{i-1})]}$$

- Finally, generate a probability by normalisation and exponenting

Benchmarks and Results

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 (Our)	81.21%

Training on CONLL 2003 Dataset

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
Weighted Average	**0.97**	**0.96**	**0.96**	**1065**

Training on CDS Sentences

Benchmarks for similar studies

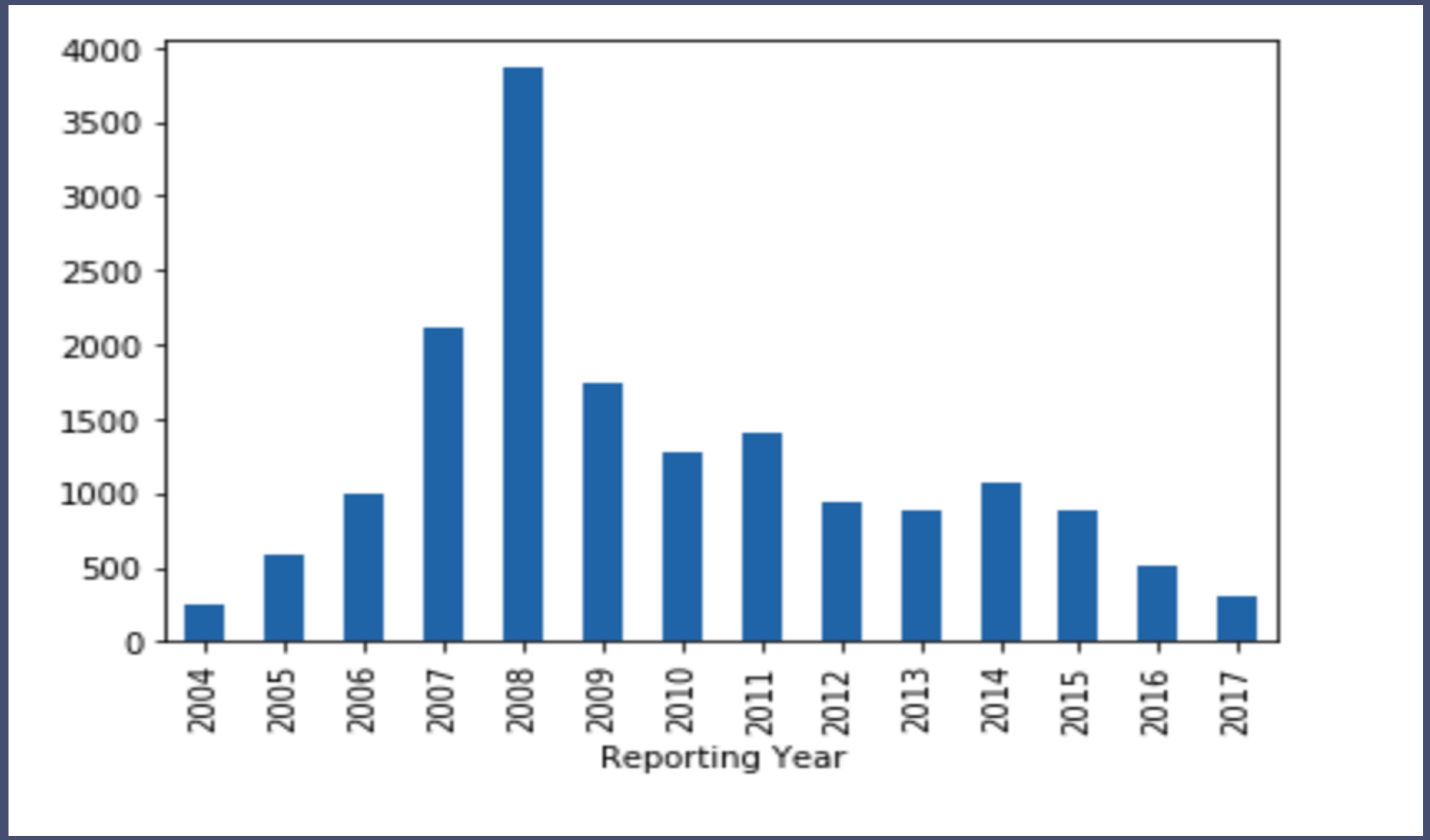
Studies Conducted	F1 Score
(Alvarado, Verspoor and Baldwin, 2013)	0.827
(Wang, Xu, Liu, Gui, and Zhou, 2015)	0.857
Bankruptcy Prediction using CRF	0.859
Our Implementation	**0.96**

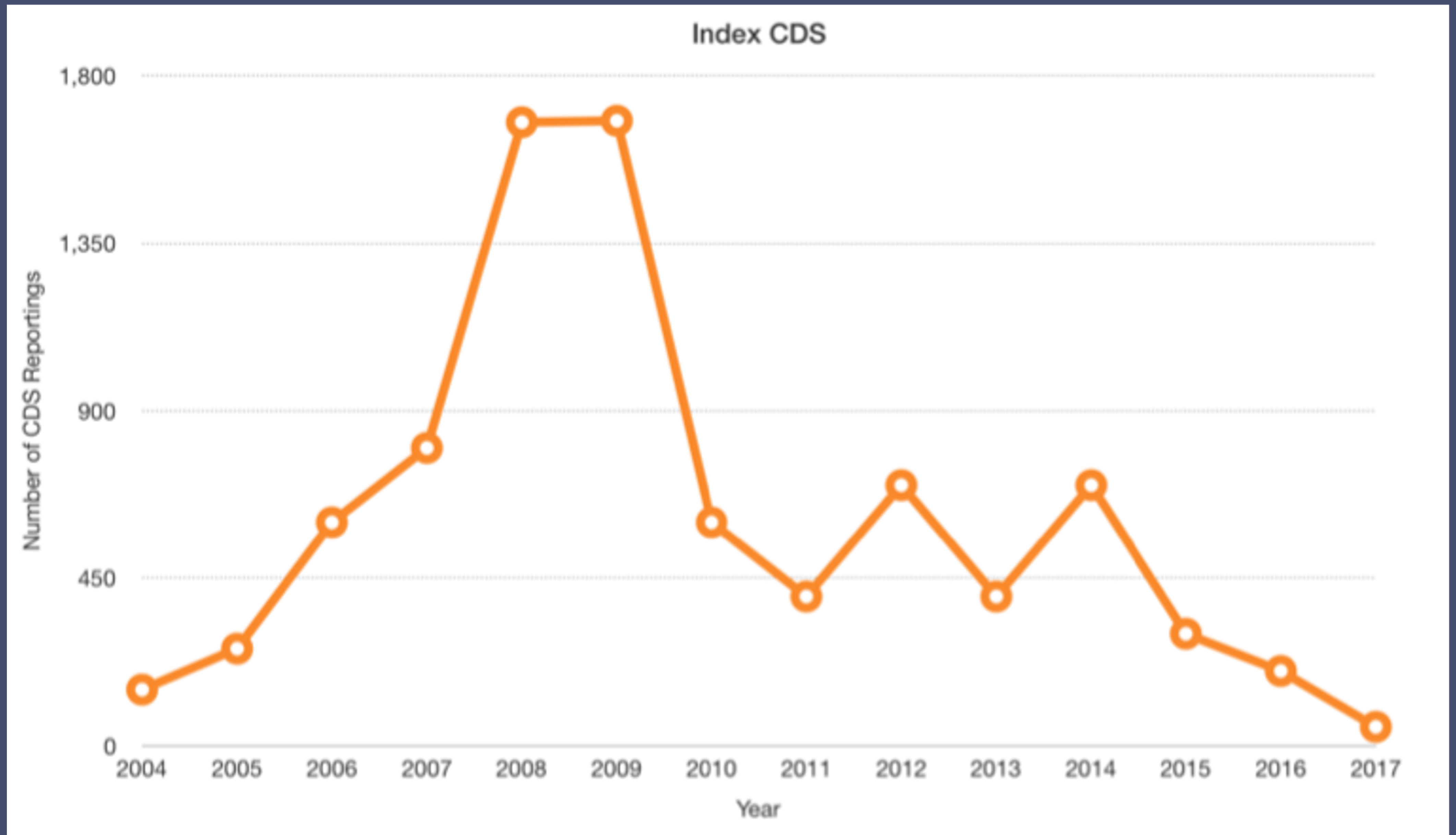
Credit risk assessment by extracting information from loan agreements

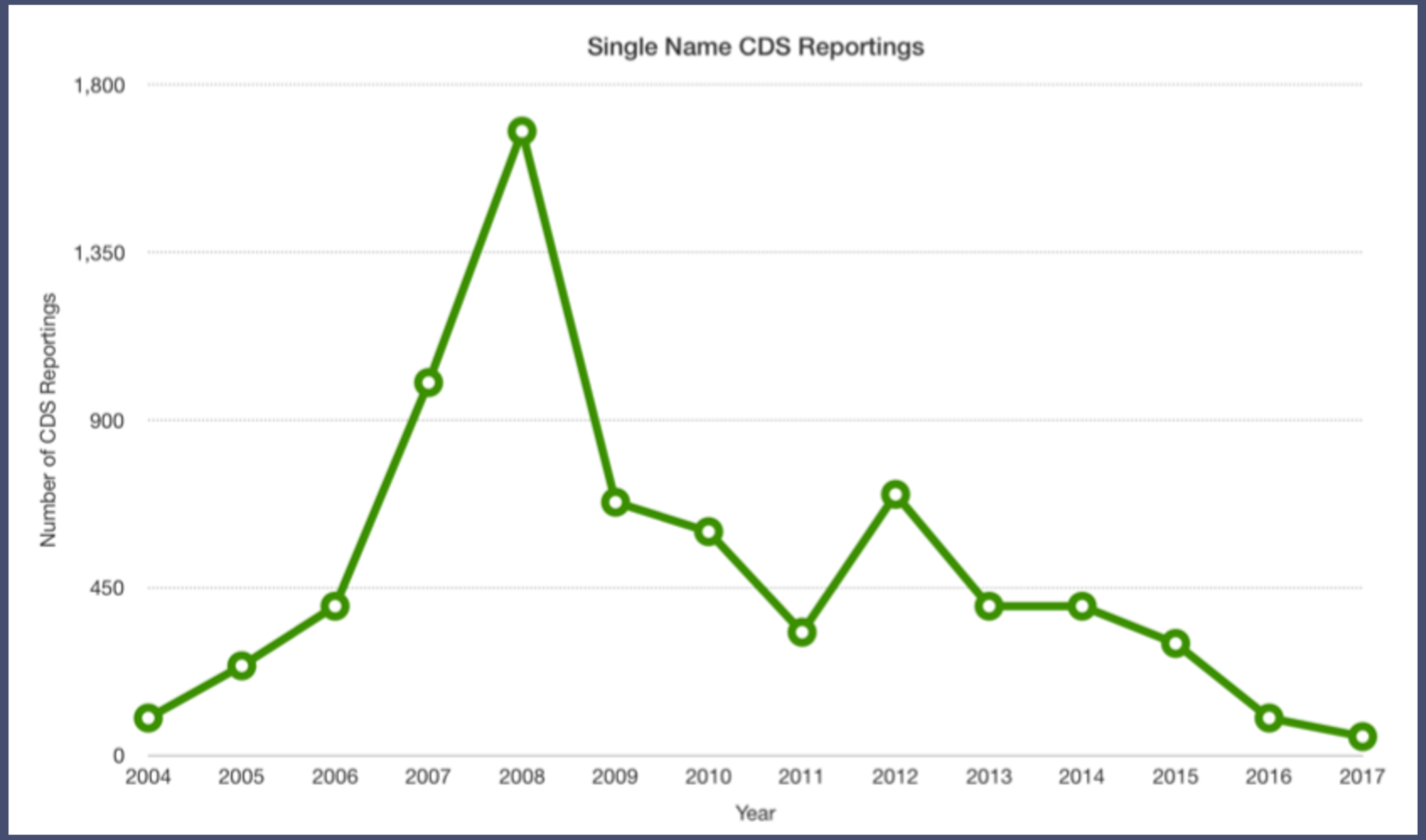
Performance benchmarks with domain-specific studies using CRF models to extract data.

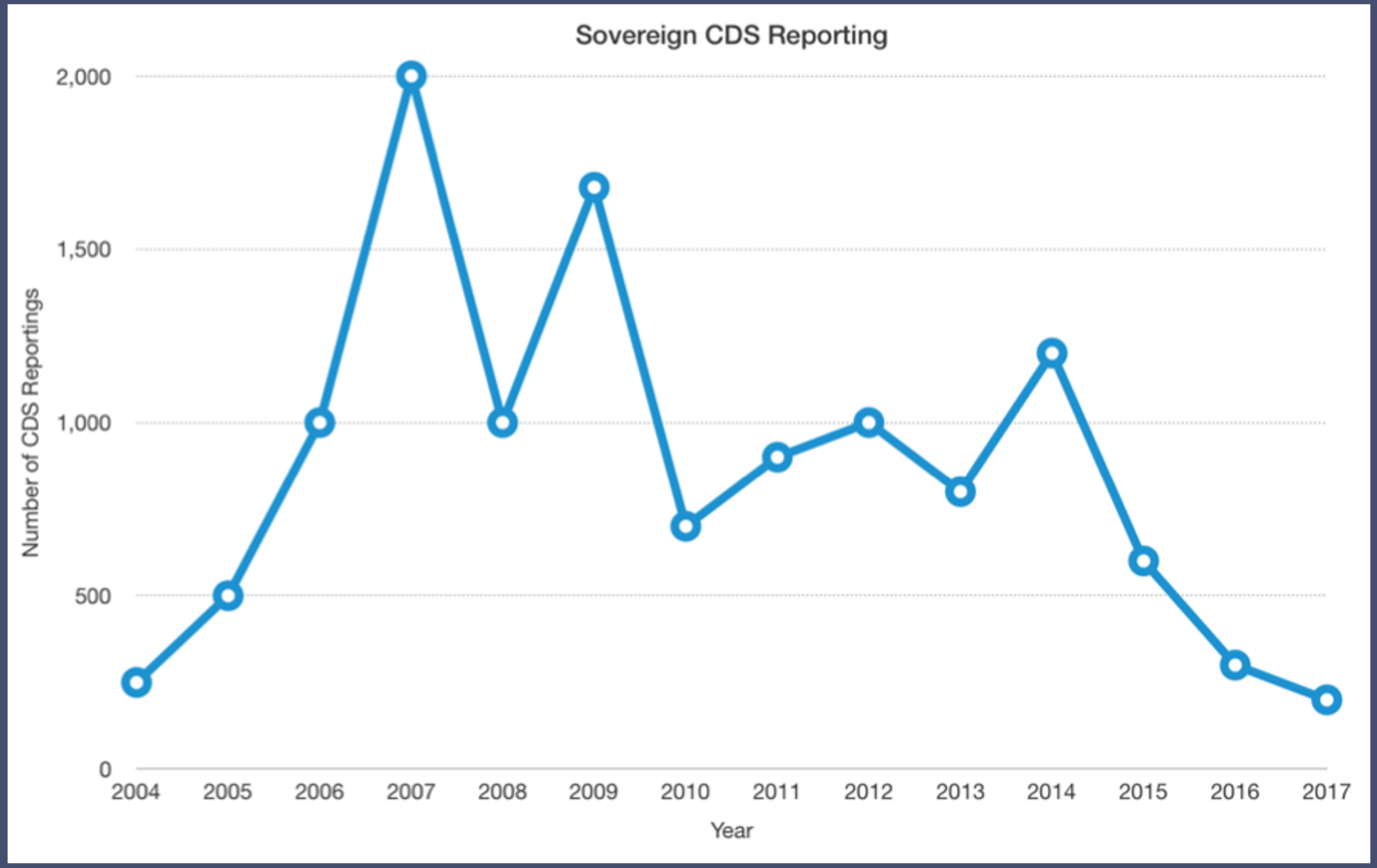
Presents a novel method to recognize named entities in financial news texts

Analysis


4000
3500
3000
2500
2000
1500
1000
500
0
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
Reporting Year


Index CDS
1,800
1,350
900
450
0
Number of CDS Reportings
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
Year


Single Name CDS Reportings
1,800
1,350
900
450
0
Number of CDS Reportings
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
Year


Sovereign CDS Reporting
2,000
1,500
1,000
500
0
Number of CDS Reportings
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
Year

Predicting Financial Health of a company

- Chance that an underlying deliverable obligation would fail to fulfill during life of the contract
- Calculate it as $q = spread/(1-R)$ where
- q is default probability (of a credit event)
- Price of a CDS is referred to as its spread and is denominated in basis points or one-hundredths of a percentage point.
- R is the assumed recovery rate and is chosen arbitrarily

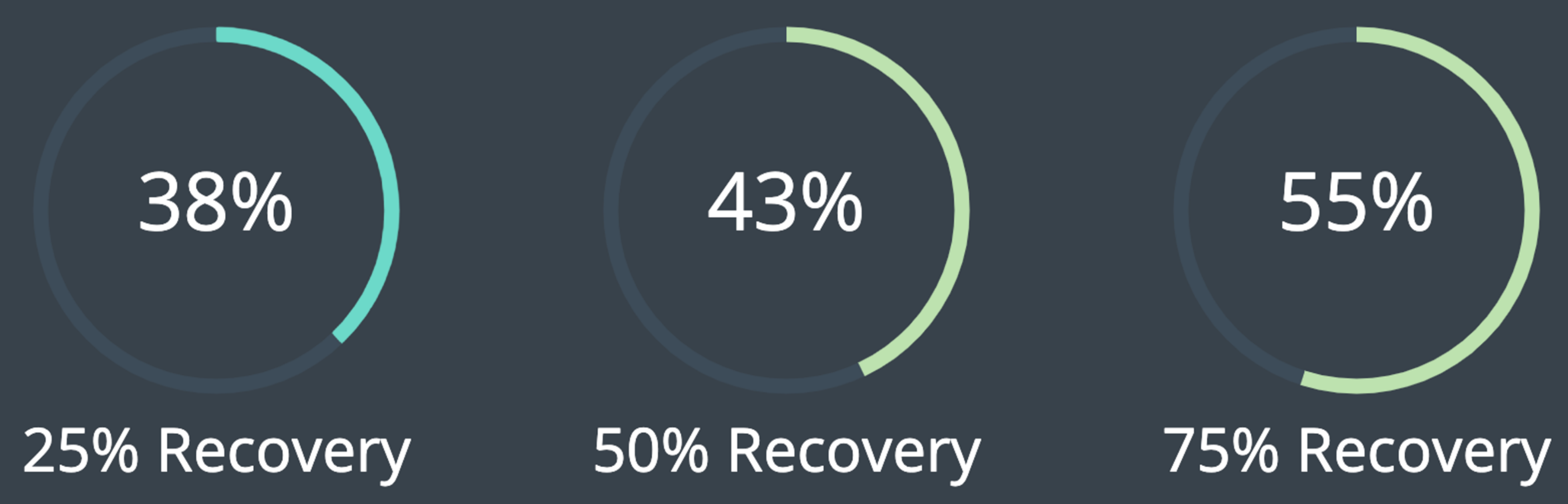
Predicting Financial Health of a company

- We get the spread for each CDS dealing from our dataset
- We calculate the default probability for each entity by taking R as 25%, 50% and 75% along with the corresponding spread.
- The probabilities are plotted on separate graphs for each value of R
- We then calculate the weighted average of all the probabilities and come up with a number to predict how well the company is doing financially

Predicting Financial Health of a company

Probability of Default of Lehman Brothers

Probability of Default of Merrill Lynch



Conclusions

- Need for a centralized database consisting of all the information regarding CDS dealings recognized.

- Both rule-based and NLP-based techniques used to extract data.

- Tools developed for preparing new datasets and expanding the current one

- Analysis done on data to answer key questions

- Developed a website combining all the models and data which is easy to use for future research

Appendix

Credit Default Swap Search

Not Secure | 0.0.0.0

Credit Default Swap Search

Raw Tables | Unified Table | Browse | Upload

Sequence Labelling

- The task of assigning a single label to each element in each sentence
- Algorithms are mostly based on probabilistic or deep learning methods
- Probabilistic methods include Conditional Random Fields and Hidden Markov Models.
- Deep Learning methods include LSTM-CRF and BiLSTM-CRF

U.N.	NNP	I-NP	I-ORG
official	NN	I-NP	O
Ekeus	NNP	I-NP	I-PER
heads	VBZ	I-VP	O
for	IN	I-PP	O
Baghdad	NNP	I-NP	I-LOC
.	.	O	O

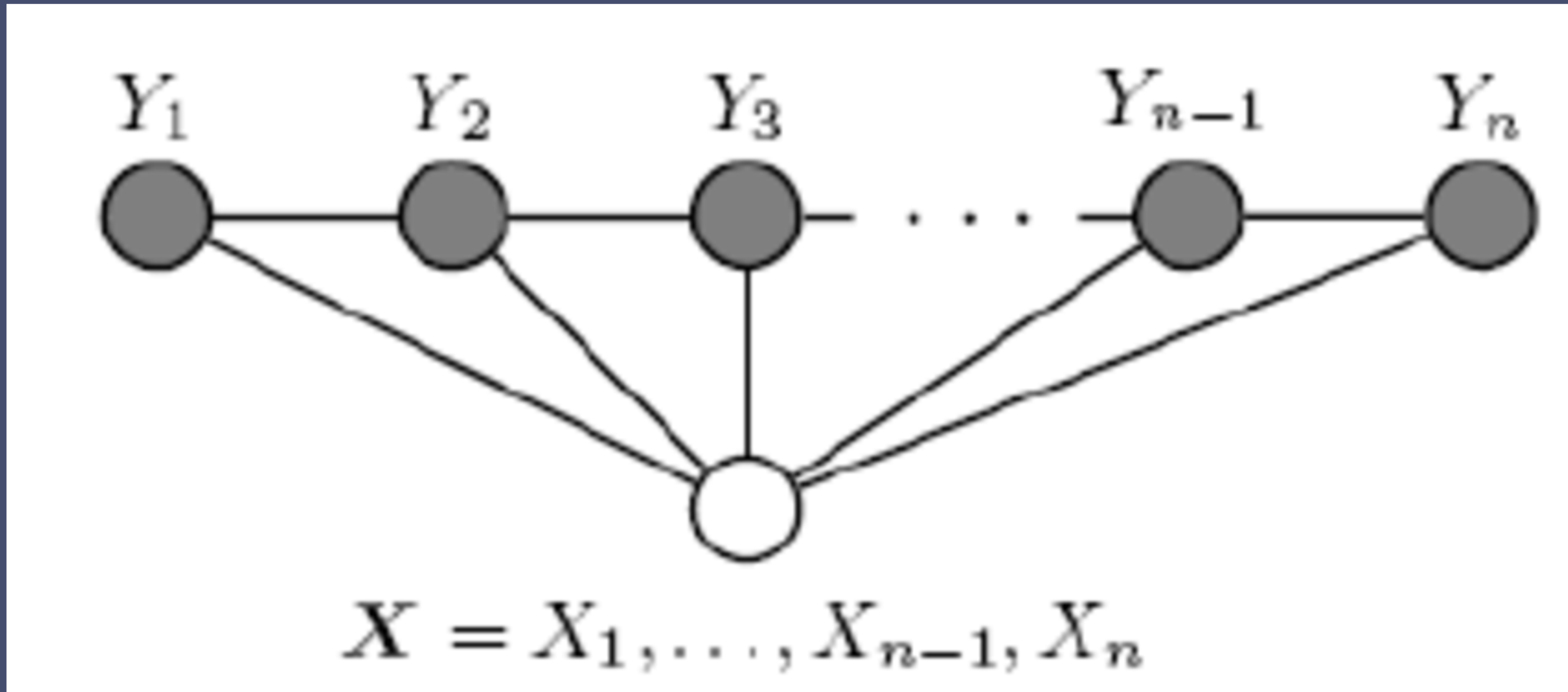
Generative vs. Discriminative Classifier

- **Discriminative models model conditional probability distribution, i.e. P(y|X)**
- **Generative models try to model a joint probability distribution, i.e., P(X,Y)**
- **Need an account for elements nearby, so conditional probability distribution needs to be considered**

U.N.	NNP	I-NP	I-ORG
official	NN	I-NP	O
Ekeus	NNP	I-NP	I-PER
heads	VBZ	I-VP	O
for	IN	I-PP	O
Baghdad	NNP	I-NP	I-LOC
.	.	O	O

Why Conditional Random Field ?

- 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)
- Support for feature functions makes CRF a better candidate over Hidden Markov



Conditional Random Field

- **Input Sequence: $X_i = [x_1, \ldots, x_n]$. target sequence of labels: $Y_i = [y_1, \ldots, y_n]$ and n is the length of the sequence**
- **U(x,y) is known as emissions scores which is essentially the score generated for a label y given the x vector at nth timestep**
- **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.**

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

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

Training the CRF Model

- **Fully labelled data is represented as $(w^{(1)}, t^{(1)}, s^{(1)})\ldots, (w^{(n)}, t^{(n)}, s^{(1)})$**

- **The objective of parameter learning is to maximize the conditional likelihood on the basis of training data.**

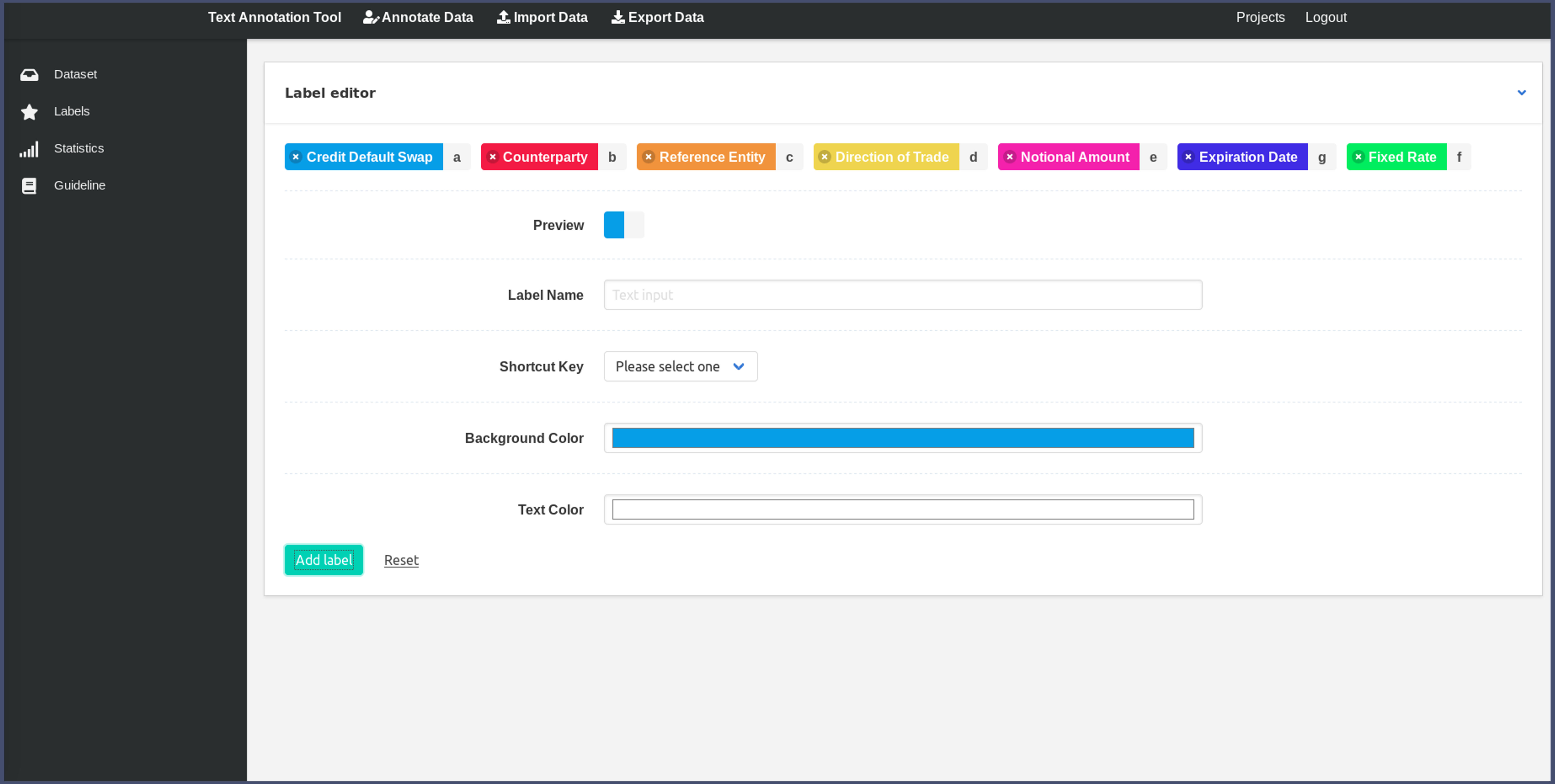
$$\sum_{j=1}^{M} logp(t^{(j)}|w^{(j)}, s^{(j)})$$

- **Conduct penalization on log-likelihood with a zero-mean Gaussian Distribution over the parameters.**

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

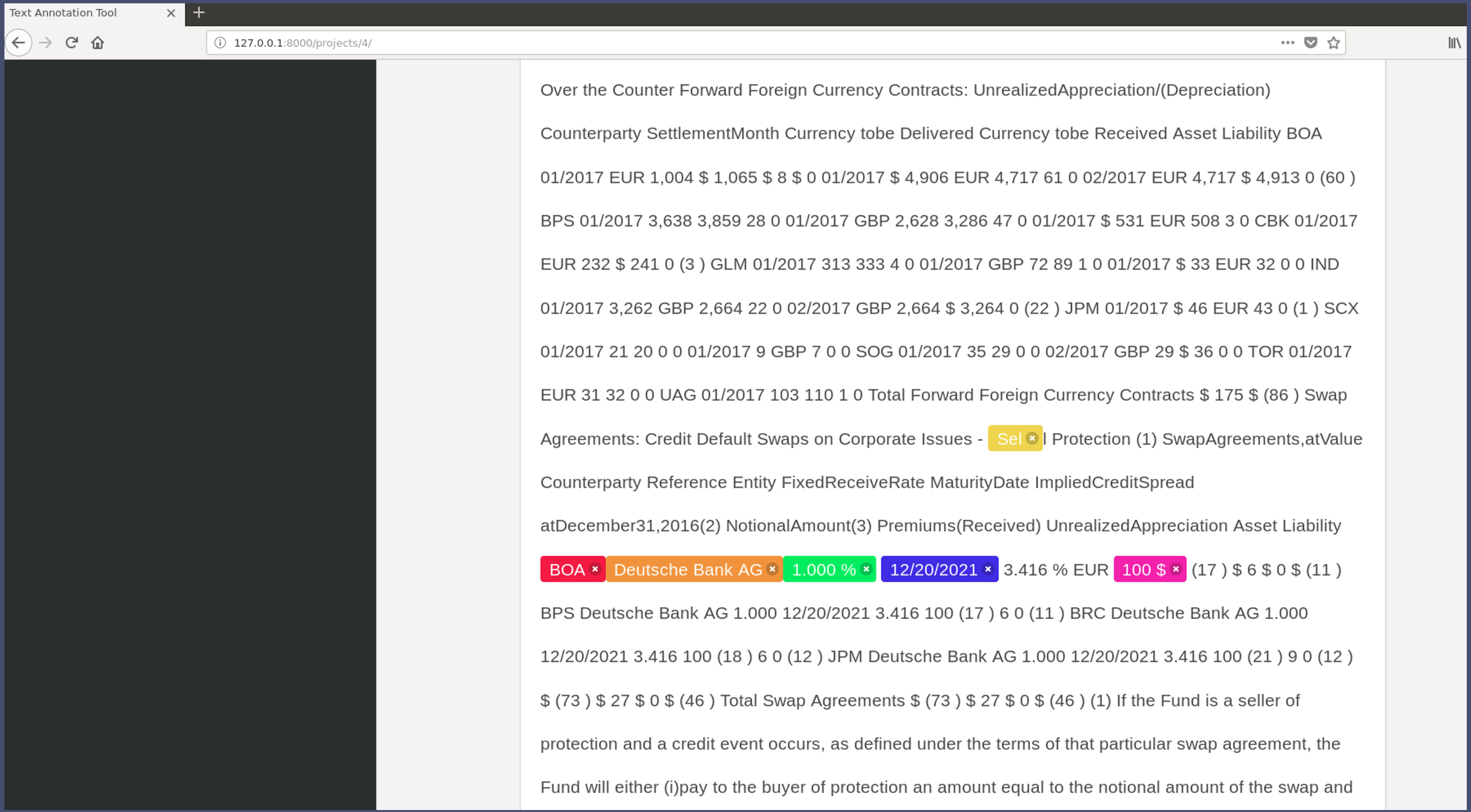
- **With the help of L-BFGS's gradient, we learn the parameters. So training the CRF model would allow us to find the optimal values of λ for the training data.**

Text Annotation Tool



Defining Labels

Text Annotation Tool



Tagging in Action