Corporate Action Summarization using Name Entity Recognition

by Sanket Sharma

Submission date: 12-Nov-2020 03:38AM (UTC+0530)
Submission ID: 1443109388
File name: Corporate_Action_Summarization_using_Name_Entity_Recognition.pdf (461.65K)
Word count: 2377
Character count: 12824

Corporate Action Summarization using Name Entity Recognition

Prof. Guide: Dr. Alok Kumar

Contributors: Vanshika Sharma, Samyak Bhagat, Sanket Sharma, Sanyukta Tanwar

ABSTRACT

Corporate actions of a company are a vital source to have an eye on its securities, cash, stock prices, etc. Mostly the release of corporate actions come in pdf documents on stock exchange websites like BSE. This project aims at summarizing documents of corporate action release of a company. The methodology includes name entity recognition. It recognizes specific keywords or phrases called entities in the text and renders them as output.

INTRODUCTION

The financial reporting of any company mostly arrives in pdf format. The annual reports and financial statements released are also published as pdf documents. Henceforth it becomes necessary to analyze the pdf format documents to gather insights about the company.

More importantly, most of the financial reporting is done around corporate actions. A corporate action is any activity that brings material change to an organization and also brings a significant impact on its stakeholders and bondholders. These events are usually acknowledged and released by the company's board of directors. Likely, the stock market trading and securities analysis also require an understanding of the corporate actions of a company as they directly impact its value.

Collectively, it becomes crucial to analyze and summarize the corporate actions of a company which are released in pdf format documents.
This project aiming at summarizing the pdf documents containing the information regarding corporate actions uses natural language processing to do the same. Specifically, *Name Entity Recognition* is used to bring out the entities in any document.

LITERATURE SURVEY

Fintech is a blend of the terms “finance” and “technology”. It alludes to any business that utilizes innovation to improve or robotize money related administrations and cycles. The term is an expansive and quickly developing industry serving the two purchasers and organizations. In today’s era, the fintech industry is rapidly growing and advancing. Fintech can likewise be considered as "any creative thoughts that improve budgetary assistance measures by proposing development courses of action as shown by different business conditions, while the musings could similarly provoke new strategies or even new associations".

In fintech, there will always be the need for technological driven business analytics. The industry is also prone to risks but only calculated ones. The computing gets us calculated decisions that are helpful in making business decisions. In the financial world, hazard to the executives is the cycle of ID, investigation, and acknowledgment or moderation of vulnerability in speculation choices. Basically, hazard the board happens when a financial specialist or asset administrator examines and endeavors to measure the potential for misfortunes in a venture, for example, an ethical danger, and afterward makes the suitable move (or inaction) given the asset's speculation goals and danger resilience.

The whole financial sector depends highly on the securities and/or cash positions. And one of the main driving factors is corporate actions. As a proprietor, the effect of corporate activity is typically estimated as far as changes to the components for example protections or potentially money positions, so corporate actions can be divided into two categories:

- Benefits: Activities that bring about an expansion in the position holder's protections or money position, without changing the fundamental security. Models incorporate extra issues, which is a Mandatory With Options Action/Event.
- Reorganizations: Activities that reshape or rebuild the underlying securities position, which at times likewise brings about a money payout. Models incorporate value rebuilds, transformations, and memberships.

PROBLEM STATEMENT

“The project aims at summarizing the pdf documents stating corporate actions released by a company to provide better insights for business analysis.”

The announced corporate action penned down in pdf documents gets released on various stock exchange sites. We will use the free text from the documents and then structure the information which can be used for further analysis and easy understanding.

SYSTEM ARCHITECTURE AND IMPLEMENTATION

Figure 1 shows the block diagram describing the process flow of the system which constitutes of four major components:

1. Data collection
2. Data pre-processing
3. NER model training
4. Entity extraction (model testing)

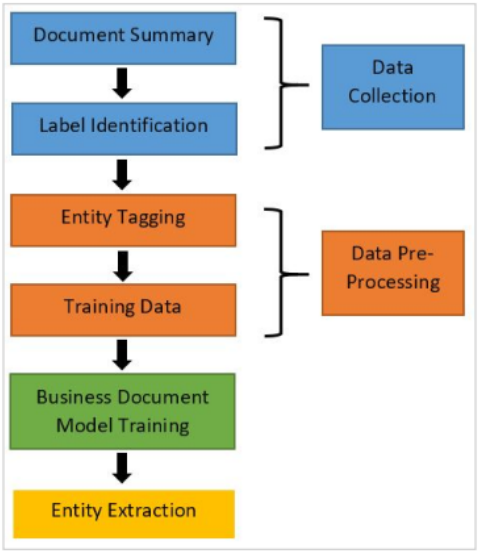

Figure 1: Process flow block diagram

1. DATA COLLECTION

The data is manually collected from the official website of BSE which is the first web platform to receive information regarding the release of any company's corporate action is BSE.The summary of the pdf documents describing corporate actions was taken and stored in a dataframe.
The entities of the corporate actions were stated and are listed below:

Corporate Action	Entity
Dividend	Offer value, Face value, Dividend percentage, Dividend decision, Announcement date, Ex-Date
Bonus/Split	Bonus require, Bonus provided, Bonus/Right Ratio, Bonus decision, Bonus/Right Ex-date
Meetings	Meeting schedule date, Meeting type, Intimation/Confirmation

2. DATA PRE-PROCESSING

The data pre-processing was done in two major steps:

2.1 ENTITY TAGGING

The entities in a text summary were tagged based on their starting and ending position.

For e.g.

Text = "UTI Asset Management Company Ltd has informed BSE that the Board of Directors of the Company at its meeting held on October 28, 2020, inter alia, has recommended a final dividend of Rs. 7/- per equity share having face value of Rs. 10/-
each for the financial year ended March 31, 2020 for approval of the Shareholders at the ensuing Annual General Meeting."

Source:https://www.bseindia.com/stockinfo/AnnPdfOpen.aspx?Pname=01FBD512-F4EF-46CC-AC8E-7B987B4F0FAF-085900.pdf

Entities	Start	End
Dividend	171	179
Offer Value	183	188
Face Value	229	235
Announcement Date	116	133
Decision	151	162

2.2 TRAINING DATA FORMATION

The NER model of Spacy accepts the training data into a specific format which is listed below:

```
train_data = [
     ("Uber blew through $1 million a week",
      {"entities":[(0, 4, 'ORG')]}),
     ("Android Pay expands to Canada",
      {"entities": [(0, 11, 'PRODUCT'), (23, 30, 'GPE')]})
]
```

The data after entity tagging was converted into the particular format.

The collected and processed data was segmented in the ratio of 8:2 wherein 80% gets into training the model and 20% in testing.

The figure below describes the system phases:

PDF document -> Entity tagging -> Entity recognition

Figure 3: Phase Diagram of tagging through docs.

3. NER MODEL TRAINING

Named entity recognition is the task of recognizing and sorting key data (entities) in text. An entity can be any word or arrangement of words that refers to something very similar. Each recognised entity is classified into category (labels).
In this project we have used spaCy's NER model to identify entities in corporate action data. spaCy is an open - source library for NLP model.

SpaCy recognizes various entities by default according to the model trained and use case. It uses pipelines to deliver linguistic attributes in a text.
In our project, we have used the "ner" pipeline.

The unique feature of spaCy is its *displaCy visualizer*. It visualizes the text based on dependency parse and named entities.

The following steps were taken:

1. Loading the model with training data and adding ner pipe and disabling the rest of other pipes to increase performability computation.
2. Adding our own labels to get recognized.
3. Shuffling and iterating over the training data to get better memorization by the model.

4. Saving the trained model.
5. Testing the model to ensure the entities in the training data are perceived effectively.

EVALUATION METRICS

After the model was implemented and trained, we needed to evaluate the performance of the trained model. In this project, we have adopted the *Confusion Matrix* to evaluate and compare our results.

Confusion Matrix is a matrix of size 2×2 for binary classification with actual values on one axis and predicted on another.

	Actual Positive	Actual Negative
Predicted Positive	True Positive	False Positive
Predicted Negative	False Negative	True Negative

Figure 4: Confusion Matrix

Accuracy is one metric for evaluating classification models. Informally, accuracy is the fraction of predictions our model got right.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \qquad \text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision can be defined as out of all the positive predicted, what percentage is truly positive.

$$Precision = \frac{TP}{TP + FP}$$

Recall can be defined as Out of the total positive, what percentage are predicted positive. It is the same as TPR (true positive rate).

$$\text{Recall} = \frac{TP}{TP + FN}$$

RESULT & ANALYSIS

The pdf document of corporate action was summarized based on entities. The output is shown via displaCy visualizer enhancing the entities needed for corporate action.

UTI Asset Management Company Ltd has informed BSE that the Board of Directors of the Company at its meeting held on October 28, 2020 ANNDATE , inter alia, has recommended DEC a final dividend DIV of Rs. 7/- OFFERV per equity share having face value of Rs. 10/-
each FACEV for the financial year ended March 31, 2020 for approval of the Shareholders at the ensuing Annual General Meeting.

Figure 5: Entity based displacy visualization

The figure below shows the confusion matrix for the entity.

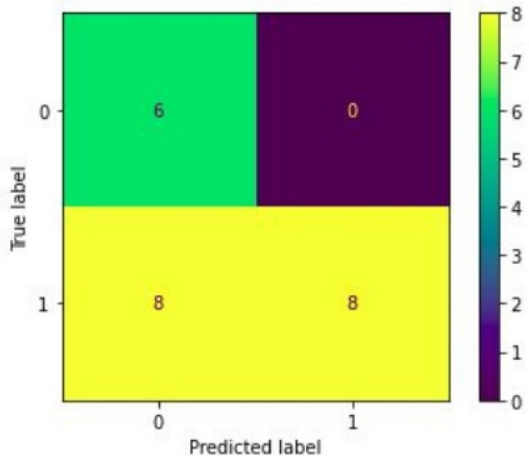

Figure 6: Confusion Matrix for entity

The evaluation metrics of each label is defined in the table below with their precision, recall, f1 score and accuracy defined.

	entity name	precision	recall	f1_score	average precision (accuracy) %
0	Meeting Scheduled (SDate)	1.0	0.777778	0.875000	97.77
1	Intimation/Confirmed (Int)	1.0	0.705882	0.827586	95.58
2	Type of Meeting (Type)	1.0	0.736842	0.848485	98.68
3	Bonus Provided (BonusProv)	0.6	0.375000	0.461538	47.50
4	Bonus Require (BonusReq)	0.5	0.500000	0.500000	45.00
5	Bonus/ Right Ratio (Ratio)	0.0	0.000000	0.000000	65.00
6	Bonus Decision (BonusDec)	1.0	0.750000	0.857143	100.00
7	Dividend (Div)	1.0	0.500000	0.666667	100.00
8	Dividend percentage (Percent)	1.0	0.400000	0.571429	67.27
9	Dividend Offer Value (OfferV)	1.0	0.500000	0.666667	86.36
10	Dividend Ex- Date (DivExDate)	0.0	0.000000	0.000000	22.72
11	Dividend Announce Date (AnnDate)	0.5	0.100000	0.166667	46.00
12	Share Face Value (FaceV)	0.0	0.000000	0.000000	45.45
13	Divided Decision (Dec)	1.0	1.000000	1.000000	100.00

Figure 7: Label wise evaluation metrics

Average evaluation matrices of the model :

Component	Total Average of each label
Precision	0.685714285714286
Recall	0.453250159713008
F1_score	0.531512924677949
average precision (accuracy) %	72.6664285714286

Inference:

1. In three of the labels (Bonus/Right Ratio, Dividend Ex-Date and Share face value) we see that despite precision and recall are null (0.00), we have an accuracy of more than 20%. This is due to True Negatives in the text.
   The testing data consisted of summaries where these entities were not present and also not predicted by the model.

2. The model is not able to predict these three entities (Bonus/Right Ratio, Dividend Ex-Date and Share face value) at all.
   This can be because:
   a. Lack of training data
   b. Dividend Ex-Date can be confused with Dividend Announced date.

The figure below shows the ROC curve for Dividend offer value.

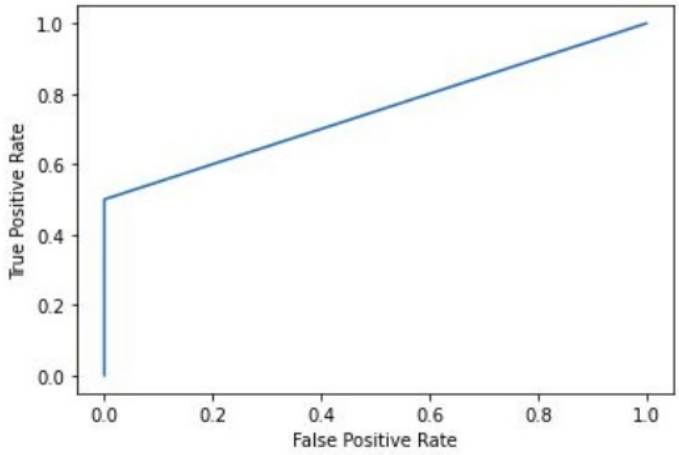

Figure 8: ROC curve for Dividend Offer Value

The figure below shows the Precision-Recall curve for Dividend offer value.

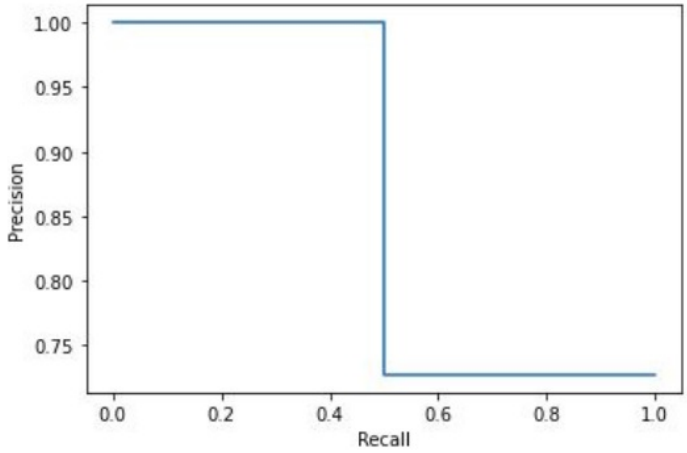

Figure 9: Precision-Recall curve for Dividend Offer Value

CONCLUSION

In this project, we recognised the entities in corporate action summaries using name entity recognition spaCy model. The trained model is capable of extracting out the entities related to different corporate actions like Offer value, Dividend percentage, Meeting scheduled date, Bonus/Right ration etc. from the provided financial documents. When tested with different evaluation matrices like confusion matrix, precision, recall and f1 score, the model shows scores as high as 0.0.82 to 0.87 for certain entities.

The output can be produced in both the formats - text and tabular.

FUTURE SCOPE

There are a lot of future scope for this project. Many things can be implemented further to increase its accuracy and precision and to make it more applicable for broder scenarios in the financial world.

1. **Training the model with more varied data** - The model sometimes suffers due to lack of variations provided in the training data through which it was trained. For example, a major part of the documents that we used in the training dataset of meetings were the intimation of a meeting of the board of directors. This was a constant type of information and the variations like voting results of AGM or EGM meetings were very less. This resulted in the model being very precise for the board meetings documents, but suffering to work through documents of AGM and EGM. More randomised training data can be used to solve this issue.

2. **Sentiment Analysis of the documents** - Every financial document can be categorised as it proving to be something which will provide a positive or negative impact on the company's finance statues. This categorisation can be done by doing a sentiment analysis of the document based on the words used and how the document is framed, deciding the sentiment whether the document is a positive or a negative news. For example, a company providing a bonus to its consumers will be considered as a positive sentiment for the company's shareholders.

3. **Improving the current model** - A big flaw of the model is that it fails in recognising some labels completely. For example, the model was unable to identify the share face value label and in any of the documents. The reason for this can be a fault while tagging the label, or maybe somehow the NER model missed to learn the label in its neural network while it was training, or due to the

lack of training data. This issue can be resolved by training the data again with more varied data and retagging the missed labels.

LEARNING OUTCOMES

During the phase of this project, we learn a lot of concepts not just related to ML and its models, but also how the finance world works.

1. **The finance and corporate actions -** We learnt about the functioning of the finance market. How the companies have to provide details of every meeting and what was concluded in it and how the information is so useful in taking decisions by shareholders was a new field that we explored. We learnt about the different corporate actions like Dividend, Meetings and Bonus and how these actions can affect the shares of a company.
2. **Data handling -** One of our big jobs in this project was to find suitable data to train and test our model. We learnt to collect and convert unstructured data to structured and meaningful data and process it as per the requirements of the model to make training and testing data.
3. **NLP modeling and training -** We learnt how to make a NLP model and train it. For this, we learnt about how different libraries like spaCy works and learnt how to make tags from a given piece of text. How to make a pipeline for the learning model and add personalised tags and how to train the prepared model with our testing data. We used sklearn to find and interpret different evaluation matrices.
4. **Evaluation Matrices -** After the completion of the project, we learnt how to evaluate a classification model and understand its performance. For this we studied confusion matrix and different parameters like precision, recall, f1 score and ROC curve.
5. **Working on collaborative platforms -** We worked on different collaborative platforms like google colab and github during the phase of this project.

AUTHOR CONTRIBUTION

Task	Samyak Bhagat	Vanshika Sharma	Sanket Sharma	Sanyukta Tanwar
Problem Formulation	✔	✔	✔	✔

Literature Review	✔	✔	✔	✔
Data Collection	✔	✔	✔	✔
Metadata Formation	✔	✔	✔	✔
Data Pre-processing		✔		
Model Formation and training	✔			✔
Entity extraction (model testing)	✔		✔	
Result preparation	✔		✔	
Result interpretation		✔		✔
Report	✔	✔	✔	✔
Poster		✔		✔

BIBLIOGRAPHY

9
https://www.investopedia.com/terms/c/corporateaction.asp
https://www.angelbroking.com/blog/corporate-actions
https://medium.com/mysuperai/what-is-named-entity-recognition-ner-and-how-can-i-use-it-2b68cf6f545d
https://en.wikipedia.org/wiki/SpaCy
https://www.investopedia.com/terms/r/riskmanagement.asp
https://www.investopedia.com/terms/a/algorithmictrading.asp#:~:text=Algorithmic%20trading%20is%20a%20process,to%20the%20market%20over%20time.

Corporate Action Summarization using Name Entity Recognition

ORIGINALITY REPORT

10%	6%	1%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.business-standard.com Internet Source	2%
2	Submitted to Asia Pacific University College of Technology and Innovation (UCTI) Student Paper	2%
3	spacy.io Internet Source	1%
4	Submitted to Leiden University Student Paper	1%
5	Submitted to University of Bedfordshire Student Paper	1%
6	Submitted to Management Development Institute Of Singapore Student Paper	1%
7	Submitted to Birkbeck College Student Paper	1%
8	widgets.economictimes.com Internet Source	1%

9	www.investopedia.com Internet Source	<1%
10	jwcn-eurasipjournals.springeropen.com Internet Source	<1%
11	Owolabi, Taoreed O., Kabiru O. Akande, and Sunday O. Olatunji. "Estimation of surface energies of hexagonal close packed metals using computational intelligence technique", Applied Soft Computing, 2015. Publication	<1%
12	www.lincolnpharma.com Internet Source	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		

Corporate Action Summarization using Name Entity Recognition