Detection and Classification of Financial Eventsfrom News Articles

PRATHAM BHOKARE[1], ANIRUDDH SONAWANE[2], SPARSH SONAWANE[3], PROF. HRUSHIKESH JOSHI[4]

[1, 2, 3] Department Of Information Technology, SCTR'S Pune Institute of Computer Technology, Pune, IN

[4] Assistant Professor, Department Of Information Technology, SCTR'S Pune Institute of Computer Technology, Pune, IN

Abstract- Latest news on financial events like Debt, stock split or mergers and acquisitions has shown significant influence on the financial market, hence it became important to keep update of these events in a timely manner for trader and investors. There are various rule-base models introduced to detect and classify the events but it requires assistance from a financial domain expert for setting rules and it also requires to updates this rules periodicaly. There are also various deep learning, cluster based and classification based approaches attempted for event detection and classification, which show significant improvement. In this work we aim to figure out efficient techniques to better detect and classify the events.

Indexed Terms- Merger and Acquisition, cluster, Rules, Deep Learning.

I. INTRODUCTION

In the finance field, the stock market and its trends are extremely volatile in nature. Investors and market analystsstudy market behavior and plan their buy or sell strategies. As the market produces large amounts of data every day, it is verydifficult for an individual to consider all scenarios and predict the future. As new articles and tweets give ideas about stock fluctuation, we are analyzing those data to give a summary of the latest feed.

In today's ever-growing financial industry, getting only an accurate insight from this data is sometimes insufficient. For applications like the stock market movement prediction, the process should be fast also. This project aims to figureout good NLP techniques requiring less supervised trainingto extract desired information from financial articles and news. Every day thousands of financial news articles and blogsare published on the internet. These articles contain valuable information and insights for business companies and individuals, especially those who are interested in the stock market. We are definitely not interested in all of the published articles,and practically it is not feasible to go through every one of them to choose the interesting articles. So to make use of this data, we need to identify and sort out valuable information in it. This is precisely what we are trying to do. The first step is collecting financial news articles from the internet and storing them in a database in an organized manner. Generally, a news article will contain nonessential introductory sentences, andwe need to filter that out. After that, we need to capture and organise the relevant information. Here we are interested only in the following financial events.

- Investment, Acquisition, Merger, Joint venture, Partnership and Lawsuit

So, in short, we need to find a method to extract relevant financial information from financial article in an organised manner.

II. OBJECTIVE

The objective of a proposed system is to help investors and traders to take appropriate financial decision. The system will help in following manner; first users will get Summarize Financial News Articles with financial events of specific company of users' interest. So, short summary of company latest financial performance also help investors and traders to save their time. Second is to detect and financial events which in appear in news article.

III. SCOPE

In this project we will try various techniques on a datasets that contains data from various sources such as financial news articles, annual financial reports and find most efficient technique for event detection and classification. Use the power of Natural Language Processing techniques to create a summary of News.

IV. SYSTEM OVERVIEW

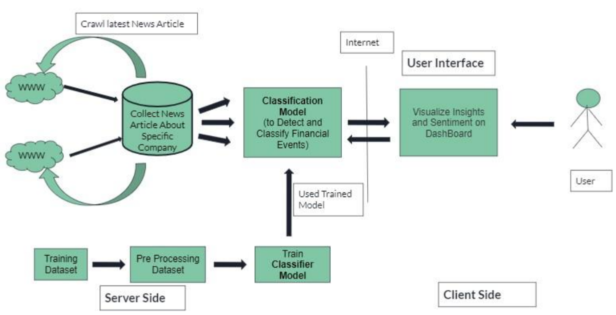

Fig. 1. System Architecture

The Working of a System -

1. Data Collection -During This Phase the data collected is from various sources available especially target data is extracted with help of web scrapper technique and data is collected continuously to get latest and updated data. the available dataset used to train and test Model
2. Building Model-the Selection of appropriate model is very necessary. it is very important to try some experimental model a and check how it performs. This very crucial step and finally model train and test using training and testing data.
3. Classifier-the build model is available so, Used this to perform classification. These models help detect financial events in the news article and also classify into the appropriate category.
4. User Interface-This is part where user is interacting with system. User will get visualization in the form bar chart and word cloud. From users able to take appropriate decision about investment.

V. RELATED WORK

A. Economic Event Detection in Company-Specific News Text[1]

This paper presents a dataset and supervised classification approach for economic event detection in English news articles. Currently, the economic domain is lacking resources and methods for data-driven supervised event detection. The Detection task is conceived as a sentence-level classification task for 11 different economic event types.

1) Dataset Used :

SentiFM Dataset which is collected from English and Dutch news articles between 2004 to 2014.

2) Algorithm Used :

Short Term Memory (LSTM) ,Recurrent Neural Network(RNN) and Linear Kernel SVM.

B. SENTiVENT: enabling supervised information extraction ofSENTIVENT

In this present an annotation scheme of company-specific events in English economic news articles. A representative corpus was crawled and annotated with an iteratively developed economic event typology with 18 Categories and 64 subcategories. This resulted in around 6200 annotated events in 288 documents. Trained model Using Concept of Transfer Learning.

1) Dataset Used :

SentiVENT Dataset which collected from English and Dutch articles between 2016 to 2017.

2) Algorithm Used :

In this paper NLP techniques BERT and RoBERT are used.

C. Financial Event Extraction Using Wikipedia-Based Weak Supervision[3]

1) Dataset Used:

Used 3 Datasets

1. Senti FMDataset (ManuallyAnnotation)
2. Random Sentence
3. Extended Wiki

2) Algorithm Used:

Binary Encoder Representational Transformer (BERT).

D. A Classification-based Approach to Economic Event Detection in Dutch News Text

This paper presented proof-of-concept experiments for a classification-based approach to detecting economic events in Dutch news text. The dataset and Algorithm are used Self annotated corpus of Dutch news articles and SVM Respectively

1) Dataset Used:

For corpus, picked news texts in which the headlines referenced at least one of the following seven Bel 20 index companies: Delhaize, Belgacom, KBC, AB InBev, Solvay, Bekaert, and Cofinimmo. The corpus utilised for the studies discussed in this study is made up of 126 articles with a total of 3,480 sentences (or 52,559 tokens).

2) Algorithm Used:

Support Vector Machine (SVM)

E. HMTC: Detecting Financial Events for Investment Decisions Based on Neural Hierarchical Multi-Label Text Classification[5]

This paper [5] proposes F-HMTC, a neural hierarchical multilabel text classification algorithm, for a financial application situation with a large number of event category labels.

1) Dataset Used :

In this paper Private self annotated datatset is used which covers all important financial events.

2) Algorithm Used:

The hierarchical clustering technique is used.

F. Banking news-events representation and classification with a novel hybrid model using DistilBERT and rule-based features[7]

This paper [7] discusses a novel hybrid approach to textclassification that incorporate a ml algorithm with DistilBERT, a pre-trained deep learning framework for natural language processing, and gives a base model fine-tuned on Indian Banking News-Events with a rule-based method that is used by separating false positives and dealing with false negatives to enhance the previous classifier's results. The main advantage is that the system may be easily fine-tuned by inserting unique rules for specific chaotic or overlapping categories that have not been adequately trained.

- Dataset Used :

In this study, they used Python-written code to scrape news from public news sources such as Bloomberg, Financial Express, Money Control, and Times of India. As a result, we have accumulated almost 10,000 occurrences of financial news stories from 2017 to 2020. The news pieces are associated with various events.

- Algorithm Used :

Hybrid model using DistilBERT and rule-based features

G. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models [8]

This study gives advice on how to work with financial sentiment. Because of the particular terminology employed in the financial sector, general-purpose models are ineffective. because it is hypothesised that pre-trained language models can help with this problem because they require fewer labelled instances and can be trained further on domain-specific corpora FinBERT, a language model based on BERT, is introduced to address NLP challenges in the financial domain. For two financial sentiment analysis datasets,

1) Dataset Used:

1. TR C2-financial dataset - they employ a financial corpus called TRC2-financial to further pre-train BERT. It is a subset of Reuters' TRC2*, which includes 1.8 million news articles produced by Reuters between 2008 and 2010.
2. Financial PhraseBank - Financial Phrase bank is made up of 4845 English sentences chosen at random from the LexisNexis database of financial news. These lines were then analysed by 16 experts with finance and business backgrounds.

2) Algorithm Used:

1. FinBERT for text classification.
2. FinBERT for regression.

VI. DATASETS PREPRATION AND FINANCIAL EVENTS DESCRIPTION

For the dataset preparation we have used sentiFm dataset[1] and sentiVent dataset[2]. We prepared a combined datasetusing the above mentioned datasets. For this combined dataset we have verified labels between the two datasets, that is, sentiFm dataset[1] and sentiVent dataset[2] after preprocessing. We observed that some labels are matching and some are not matching. For the matching labels we have directly merged them under the combined dataset. And for the non matching labels we have used the following similarity techniques to find appropriate labels –

1) Cosine similarity:
2) Pretrained Sentence similarity Model:
3) Manually checked:

A. Finally we have obtained 11 Financial Events in combined Dataset as follows –

1) Dividend: It represents financial news which contain dividend related text. Dividend part of a company's profits thatis paid to the people who own shares in it (shareholders).
2) Rating and Recommendation: Events on the projected price level of a security. We include announcements, fore- casts, price raised, reduced, or maintained and also shows A recommendation to purchase the security from an analyst. As event mentions, we include rating announcements, forecasts, performance, buy/sell/hold advice, and rating up- grades/downgrades/maintained.
3) Partnership and Joint Venture: it represents news related to Agreement and Deal between two or more entity.
4) Share Repurchase: Share Repurchase events by a company including announcements and forecasts of share repurchases.
5) Financial Results: Financial benefits that are realized when the amount of revenue exceeds expenses. We include declarations and forecasts of profit, sales volumne, positive and negative (losses) profit, lower than, higher than, as expected, increased, decreased, and stable profits.
6) Investment and Funding: Investment and Funding events including investment done amd fund raised by company.
7) unknown: unknown events includes all news which isa financial but belongs other category.
8) Debt: Event mentions pertaining to company debt and debt ratios. We include debt announcements, forecasts, in- creases, reductions, and restructuring.
9) Merger and Acquisition: Mergers and acquisitions refers to the consolidation of companies or assets involving at least two companies. Weinclude announcements, forecasts, and cancellations of a merger/acquisition.
10) Securities Turnover: The number and frequency of securities traded over a certain period. We include declaration and prediction of turnover figures, increased, decreased, stable, worse than, better than, and as expected turnover.
11) litigation: it is event include company law related news

VII. DATASET AND IT'S DESCRIPTION

A. SentiFM dataset[1]

This dataset is prepared from the English and Dutch news articles which contains a feature as financial text from these news articles and labels which show financial events as the output, also contains 4098 rows and 11 unique labels as sales volume, target price, rating, buy purchase, dividend, turnover, debt, profit, merger and acquisition, and Quarterly results.

Sr No.	Label	Count
1	Debt	55
2	Dividend	191
3	Financial_Results	1183
4	Investment and Funding	64
5	Litigation	52
6	Merger_and_Acquisition	348
7	Partnership_and_JointVenture	105
8	Rating_and_Recommendation	275
9	Securities_Turnover	155
10	ShareRePurchase	41
11	Unknown Event	2469
	Total	4968

Fig. 2. SentiFM Dataset

B. SentiVent dataset [2]

This dataset is also prepared from the English and Dutch news articles which contains a feature as financial text from these news articles and labels which show financial events as the output, but the difference is that, SentiFm is single class dataset and SentVent is multiclass dataset which contains 6156 rows and 481 unique labels. Thus, we preprocess the dataset by renaming the labels and reducing the count to 18 labels

Sr no	Label	Count
1	"['unknown']"	3422
2	['Financial_Results', 'Securities_Turnover']	47
3	['Financial_Results']	842
4	['Securities_Turnover']	321
5	['Rating_and_Recommendation']	102
6	['Investment_and_Funding']	618
7	['Partnership_and_JointVenture','Investment_and_Funding']	22
8	"['Partnership_and_JointVenture']	120
9	['Debt']	49
10	['Dividend]	60
11	[litigation']	304

Fig. 3. SentiVent Datset

12	['Financial_Results', 'Investment_and_Funding']	80
13	['litigation', 'Investment_and_Funding']	35
13	['Financial_Results', 'litigation']	17
15	['litigation', 'Partnership_and_JointVenture']	17
16	['Financial_Results', 'Rating_and_Recommendation']	19
17	['Rating_and_Recommendation','Securities_Turnover']	25
18	['Merger_and_Acquisition']	56
	Total	6156

Fig. 4. SentiVent Datset

C. Combined dataset

This is the final dataset which is a combination of SentiFm Dataset and SentiVent Dataset. This dataset contains 10818 rows and 11 unique labels. Further we have used this dataset for training and testing purposes for out model.

Sr No.	Label	Count
1	Debt	728
2	Dividend	251
3	Financial_Results	2410
4	Investment and Funding	64
5	Litigation	52
6	Merger_and_Acquisition	994
7	Partnership_and_JointVenture	251
8	Rating_and_Recommendation	64
9	Securities_Turnover	52
10	Share_RePurchase	105
11	Unknown Event	5891
	total	10818

VIII. METHODOLOGY

We split our problem into sub-parts, we can make use of other available financial datasets. This enables to cover more 'event' types and helps to increase the extraction accuracy. We adopted the second approach and sub-divided the problem as follows.

1) *Part 1: Collection of financial news articles:*
2) *Part 2:Detection of financial events from news articles:*
3) *Part 3: Classification of financial events into different categories:*
4) *Part 4: Extraction information related to each event.: An overview of the sub-divisions in our problem is given in the following diagram.*

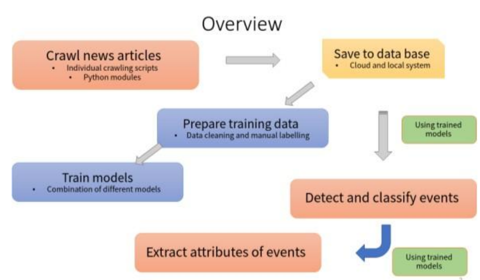

Fig. 6. Overview

A. Financial Event Detection

Once we have collected all the financial news then at the sentence level we have checked whether the sentence is financial or not. For performing this task we define the binary classifier which classifies the sentence into 0 or 1 then all these sentences are given to financial event classification.

1) *Logistic Classifier:* In this experiment, we trained a Logistic regression classifier for binary classification. Regarding the training dataset, the combined dataset has more samples and more event types,. So we selected all event types from the combined dataset, relabelled all into 'event' and added them to our labelled dataset. Here the sentences are only labelled whether they belong to 'event' or 'non-event'; no detailed labellings are needed.

B. Financial Event Classification

Once a sentence is detected as a financial sentence, forclassifying these sentences into different financial events. Following are the approaches used for classification-

1) *Machine Learning Approach:* In this approach we have worked upon five different models which are Logistic regression, Linear SVC, Naive Bayes, XGBoost and Random Forest. From all these models the Linear SVC has given Better Result. Our main aim behind using ml models is that they are lightweight and simple as compared to DL Models.

1. Logistic regression classifier The logistic regression model is trained for classification with our training dataset. Both the "one versus one" (ovo) and "one versus all" (ova) methods are tried. The trained model is validated on the validation dataset, and the performance of the 'ovo' model was relatively better than 'ova'. The classification report is given below.

	precision	recall	f1-score	support
0	0.83	0.11	0.20	87
1	0.94	0.31	0.47	100
2	0.00	0.00	0.00	43
3	0.00	0.00	0.00	33
4	0.73	0.63	0.67	491
5	0.00	0.00	0.00	35
6	0.67	0.95	0.78	1191
7	0.00	0.00	0.00	21
8	1.00	0.04	0.08	92
9	0.00	0.00	0.00	33
10	1.00	0.08	0.15	38
accuracy			0.69	2164
macro avg	0.47	0.19	0.21	2164
weighted avg	0.67	0.69	0.62	2164

Fig. 7. Logistic Regression Classification Report

2. Linear SVC regression classifierear SVC is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVC classifier. the model is trained using training dataset. The trained model is validated on the validation dataset. The classification report is given below

	precision	recall	f1-score	support
0	0.84	0.76	0.80	87
1	0.83	0.74	0.78	100
2	0.95	0.81	0.88	43
3	0.85	0.52	0.64	33
4	0.79	0.80	0.79	491
5	0.90	0.77	0.83	35
6	0.83	0.89	0.86	1191
7	0.76	0.90	0.83	21
8	0.78	0.54	0.64	92
9	0.81	0.52	0.63	33
10	0.65	0.34	0.45	38
accuracy			0.82	2164
macro avg	0.82	0.69	0.74	2164
weighted avg	0.82	0.82	0.81	2164

Fig. 8. Linear SVC Classification Report

3. Naive Bayes classifier

Na¨ıve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. the model is trained using training dataset. The trained model is validated on the validation dataset. The classification report is given below

	precision	recall	f1-score	support
0	0.81	0.44	0.57	87
1	0.91	0.51	0.65	100
2	0.92	0.26	0.40	43
3	1.00	0.03	0.06	33
4	0.78	0.75	0.76	491
5	1.00	0.09	0.16	35
6	0.73	0.94	0.82	1191
7	0.67	0.19	0.30	21
8	0.88	0.23	0.36	92
9	1.00	0.15	0.26	33
10	0.83	0.13	0.23	38
accuracy			0.75	2164
macro avg	0.87	0.34	0.42	2164
weighted avg	0.78	0.75	0.72	2164

Fig. 9. Naive Bayes Classification Report

4. XGBoost classifier

it is an optimized distributed gradient boosting library designed for efficient and scalable training of machine learning models. It is an ensemble learning method that combines the predictions of multiple weak models to produce a stronger prediction. XGBoost stands for “Extreme Gradient Boosting” and it has become one of the most popular and widely used machine learning algorithms due to its ability to handle large datasets and its ability to achieve state-of-the-art performance in many machine learning tasks such as classification and regression. the model is trained using training dataset. The trained model is validated on the validation dataset. The classification report is given below

	precision	recall	f1-score	support
0	0.74	0.56	0.64	87
1	0.89	0.56	0.69	100
2	0.89	0.40	0.55	43
3	0.56	0.15	0.24	33
4	0.78	0.65	0.71	491
5	1.00	0.06	0.11	35
6	0.72	0.91	0.80	1191
7	0.75	0.71	0.73	21
8	0.75	0.23	0.35	92
9	0.60	0.45	0.52	33
10	0.62	0.21	0.31	38
accuracy			0.74	2164
macro avg	0.75	0.45	0.51	2164
weighted avg	0.75	0.74	0.71	2164

Fig. 10. XGBoost Classification Report

5. Random Forest classifier

The random forest is a classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree. the model is trained using training dataset. The trained model is validated on the validation dataset. The classification report is given below

	precision	recall	f1-score	support
0	0.91	0.70	0.79	87
1	0.96	0.70	0.81	100
2	0.92	0.81	0.86	43
3	0.95	0.64	0.76	33
4	0.76	0.82	0.79	491
5	0.86	0.71	0.78	35
6	0.82	0.90	0.85	1191
7	0.93	0.62	0.74	21
8	0.76	0.51	0.61	92
9	0.77	0.30	0.43	33
10	0.75	0.24	0.36	38
accuracy			0.81	2164
macro avg	0.85	0.63	0.71	2164
weighted avg	0.82	0.81	0.81	2164

Fig. 11. Random Forest Classification Report

2) Transfer Learning Approach:

1) *Bert Model:* BERT is a transformers model pretrained on a large corpus of English data in a self-supervised fashion. This means it was pretrained on the raw texts only, with no humans labeling them in any way (which is why it can use lots of publicly available data) with an automatic process to generate inputs and labels from those texts. More precisely, it was pretrained with two objectives:

1. Masked language modeling (MLM).
2. Next sentence prediction (NSP).

We fine tuned this model and trained on training dataset and tested on validation dataset. The model is performing betterthan ml models. The classification report is given below-

```
***** Running Evaluation *****
  Num examples = 2164
  Batch size = 8
[271/271 00:11]
<class 'transformers.trainer_utils.EvalPrediction'>
```

	precision	recall	f1-score	support
0	0.87	0.90	0.88	87
1	0.88	0.92	0.90	100
2	0.91	0.93	0.92	43
3	0.75	0.73	0.74	33
4	0.81	0.87	0.84	491
5	1.00	0.74	0.85	35
6	0.90	0.89	0.89	1191
7	0.72	0.86	0.78	21
8	0.82	0.73	0.77	92
9	0.84	0.64	0.72	33
10	0.73	0.50	0.59	38
accuracy			0.87	2164
macro avg	0.84	0.79	0.81	2164
weighted avg	0.87	0.87	0.86	2164

```
{'eval_loss': 0.9455037117004395,
 'eval_accuracy': 0.8650646950092421,
 'eval_precision': 0.8650646950092421,
 'eval_recall': 0.8650646950092421,
 'eval_f1': 0.8650646950092421,
 'eval_runtime': 11.6334,
 'eval_samples_per_second': 186.017,
 'eval_steps_per_second': 23.295,
 'epoch': 6.0}
```

Fig. 12. BERT Classification Report

2) *Financial Text Classification:* Model determines the financial sentiment of the given text. Given the unbalanced distribution of the class labels, the weights were adjusted to pay attention to the less sampled labels which should increase overall performance. The COVID dataset was added in orderto enrich the model, given most models have not been trained on the impact of Covid-19 on earnings or markets. When We fine tuned this model and trained on training dataset and tested on validation dataset. the model give better result than bert model .the classification report is given below-

```
***** Running Prediction *****
  Num examples = 2164
  Batch size = 8
<class 'transformers.trainer_utils.EvalPrediction'>
              precision    recall  f1-score   support

           0       0.83      0.90      0.86        87
           1       0.88      0.90      0.89       100
           2       0.91      0.93      0.92        43
           3       0.81      0.79      0.80        33
           4       0.80      0.87      0.83       491
           5       0.91      0.86      0.88        35
           6       0.90      0.85      0.88      1191
           7       0.76      0.90      0.83        21
           8       0.69      0.79      0.74        92
           9       0.83      0.73      0.77        33
          10       0.62      0.55      0.58        38

    accuracy                           0.85      2164
   macro avg       0.81      0.82      0.82      2164
weighted avg       0.86      0.85      0.85      2164
```

Fig. 13. Financial Text Model Classification Report

3) *Distillbert Model:* DistilBERT is a transformers model, smaller and faster than BERT, which was pretrained on the same corpus in a self-supervised fashion, using the BERTbase model as a teacher. This means it was pretrained on the raw texts only, with no humans labelling them in anyway (which is why it can use lots of publicly available data) with an automatic process to generate inputs and labels from those texts using the BERT base model. More precisely, itwas pretrained with three objectives:

1. Distillation loss.
2. Masked language modeling (MLM).
3. Cosine embedding loss.

This is final model we tried finetined and trained on training dataset and tested on validation dataset. the model is performing better than rest of the models. the classificationreport is given below-

```
***** Running Evaluation *****
  Num examples = 2164
  Batch size = 8
[271/271 00:11]
<class 'transformers.trainer_utils.EvalPrediction'>
              precision    recall  f1-score   support

           0       0.83      0.89      0.86        87
           1       0.83      0.91      0.87       100
           2       0.91      0.93      0.92        43
           3       0.88      0.88      0.88        33
           4       0.81      0.85      0.83       491
           5       0.97      0.86      0.91        35
           6       0.89      0.87      0.88      1191
           7       0.74      0.95      0.83        21
           8       0.77      0.76      0.77        92
           9       0.82      0.70      0.75        33
          10       0.75      0.47      0.58        38

    accuracy                           0.86      2164
   macro avg       0.84      0.82      0.83      2164
weighted avg       0.86      0.86      0.86      2164

{'eval_loss': 1.180109262466307,
 'eval_accuracy': 0.8585951940850277,
 'eval_precision': 0.8585951940850277,
 'eval_recall': 0.8585951940850277,
 'eval_f1': 0.8585951940850277,
 'eval_runtime': 11.6518,
 'eval_samples_per_second': 185.723,
 'eval_steps_per_second': 23.258,
 'epoch': 8.0}
```

Fig. 14. Distillbert Model Classification Report

IX. EXPERIMENTAL RESULTS

Following are training shown in table 1,2 and testing shown in table 3 results of below given models to understand the result or perfomance we need to look following two term -

1) *Macro Average:* calculates the F1 separated by class butnot using weights for the aggregation.
2) *Weighted Average:* calculates the F1 score for each class independently but when it adds them together uses a weight that depends on the number of true labels of each class.

A. *Results of Machine Learning Models*

We have tested the above ML models on validation dataset, as results show that Linear SVC gives better results than the rest of the models.

Vectorizer	Model	Macro Avg	Weighted Avg
TFIDF	Random Forest	0.71	0.80
	LinearSVC	0.74	0.81
	MultinomialNB	0.21	0.62
	Logistic Regression	0.42	0.72
	XGBoost	0.51	0.71

Fig. 15. Results of Machine Learning Approach

B. *Results of Transfer Learning Models*

From the transfer learning technique We have tested the above models on validation dataset, as results show that Linear SVC gives better result than the rest of the models.

Model	Macro Avg	Weighted Avg
Bert-base-uncased	0.81	0.86
Financial-text-classification	**0.82**	**0.85**
Distillbert	0.83	0.86

Fig. 16. Results of Transfer Learning Approach

Conclusion From above we conclude that from the ML models the Linear SVC performs better and from transfer learning techniques the Distillbert model is performing well.

C. *Testing Results*

The Following are the results which shows performance of model on a various testing dataset as given below-

Model	Dataset	Macro Avg(Top Label)	WeightedAvg(Top Label)
	Test -01(Eng and Dutch News)2194	0.74	0.81
LinearSVC	Test -02(Indian News)109	0.44	0.61
	Test -03(Indian News)202	0.42	0.62
	Test -04(Indian News)110	0.61	0.61
	Test -01(Eng and Dutch News)2194	0.83	0.86
Distillbert	Test -02Indian News)109	0.49	0.70
	Test -03Indian News)202	0.46	0.67
	Test -4(Indian News)110	0.66	0.66

Fig. 17. Testing Results

CONCLUSION

This research paper focused on financial news articles, tweets, posts to detect and classify financial events from these contents. sparsity of financial data has always been major problem in development of successful classification model. Hence, we have tried to accumulate various event detection and classification schemes from the last decade so that new researchers can glean ideas from them and employ new robust models to detect and classify events. we systematically selected various research articles and presented a brief analysis of technique used for classification, their performance metrics, and performance analysis. We found that most of the models follow a general approach with data collection, preprocessing, processing, classification. We focused only on financial news articles in this work, and we can extend the work to include more financial datasets like blogs, tweets, and financial documents. Also, we can include more financial event types other than the events discussedin this study. For the event extraction part, we have done only the major six event types out of the ten event types we successfully classified. We can include the rest of the event types by creating a detailed labeled training dataset for each.

We will also try deep learning techniques to see if it canprovide better result. To be specific:

- In the data collection phase, we noted their used datasets, collection process, and the regions from where they collected those data. Different researchers used different languages (e.g., English, dutch, etc.) to detect and classifyevents.
- In the preprocessing steps, data have been cleaned for processing. Various features have been collected and later sent for classification step in the processing phase.
- In the classification step, the articles discuss financial events such as debt, sales volume, revenue, merger and acquisition, and many others.
- The key performance metrics were precision, recall, ac- curacy, F-score, specificity, sensitivity, TP, TN, FP, FN, ROC, AUC, confusion matrix, etc.

REFERENCES

[1] gillesm.jacobs,els.lefever,veronique.hoste, ”Economic Event Detection in Company-Specific News Text”, Proceedings of the First Workshop on Economics and Natural Language Processing, pages 1–10 Melbourne, Australia, July 20, 2018. c 2018 Association for Computational Linguis-tics.

[2] Gills Jacob,Veronique Hoste, ”SENTiVENT: enabling supervised infor- mation extraction of companyspecific events in economic and financial news”, Lauanguage and Translatin team,Ghent University 8 october2021 belguim

[3] Liat Ein-Dor, Ariel Gera, Orith Toledo-Ronen, Alon Halfon, Benjamin Sznajder, Lena Dankin, Yonatan Bilu, Yoav Katz and Noam Slonim, ”Financial Event Extraction Using WikipediaBased Weak Supervi-sion”,roceedings of the Second Workshop on Economics and Natural Language Processing, pages 10–15 Hong Kong, November 4.c 2019 Association for Computational Linguistics

[4] Els Lefever and Veronique Hoste ”A Classification-based Approach to Economic Event Detection in Dutch News Text”, T3 Language and Translation Technology Team Ghent University GrootBrittannielaan 45, 9000 Ghent, Belgium ¨ els.lefever, veronique.hoste

[5] Xin Liang1,2, Dawei Cheng3,4Fangzhou Yang4,Yifeng Luo1,2Weining Qian 1,2, Aoying Zhou1, ”F-HMTC: Detecting Financial Events for Investment Decisions Based on Neural Hierarchical Multi-Label Text Classification”,

Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20) Special Track on AIin FinTech.

[6] Varun Dograa,Sahil Vermab,*,Aman Singha,Kavitab,M N Talibc

[7] ,Mamoona Humayun ”Banking news-events representation and classi- fication with a novel hybrid model using DistilBERT and rule-based features Turkish Journal of Computer and Mathematics Education Vol.12 No.10 (2021), 3039-3054

[8] FinBERT: Financial Sentiment Analysis with Pre-trained Language ModelsDogu Tan Araci dogu.araci@student.uva.nl University of Amsterdam Amsterdam, The Netherlands