

# **INTRODUCTION**

## **1.1 Stock Market Prediction**

Stock market investment is a process of predicting the future values of stocks. The investment decisions are mainly driven by the market information available to investor. These predictions help any significant market to yield profit.

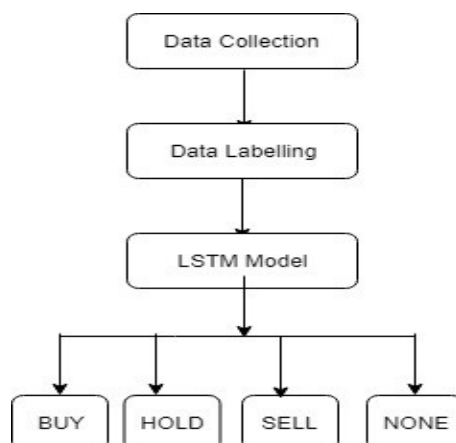
## **1.2 More about the Project**

Predicting the Stock Market has been the bane and goal of investors since its existence. Entire day companies rise and fall everyday based on the behaviour of the market. Despite its prevalence, it is quite secretive and deceptive art. The chief goal of this project is to enhance the understanding of stock market prediction. The better understanding includes how the market moves, what investors think about every ups and downs of stocks. In this project the main motive was to predict the prediction of stocks through social network analysis. Basically social network analysis means we looked out for comments, messages or reviews on the website moneycontrol.com where many investors related to particular market predict the stock movement through commenting their views. These comments highlight three basic prediction either to “Sell” or “Buy” or “Hold”. The project will make no attempt to deciding how much money to allocate to each prediction.

## **Objective**

To predict the stock movements using LSTM network.

## **Proposed Model**



# DATASETS AND FEATURES

## 2.1 Data-set and Preparation

The Data-set was collected from moneycontrol.com by using one handy software called Octo-parse which has a feature of web scrapping. The four different company's stocks were scrapped Airtel, PC Jeweller, Bajaj Finance and Tata Steel.

### Columns Extracted:-

user\_name, num\_messages, stocks\_predicted, user\_level, followers, messages, time, ratings and offensive.

The main feature that was used in this project was 'messages'.

## 2.2 Process of Labeling

### Two ways of labeling:-

#### 1. By coding in Python

```
In [2]: import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk import pos_tag
import pandas as pd
df=pd.read_excel('Airtel_MARKED.xlsx')
messages=df['message']
li=['buy','hold','sell']

for i in range(0,380):
    message=str(messages.iloc[i:i+1])
    tokened=list(word_tokenize(message))
    for words in tokened:
        if words.lower()=='buy' or words.lower()=='sell' or words.lower()=='hold':
            print(str(tokened))

In [17]:
```

```
[ '0', 'hold', 'your', 'holdings', 'it', 'may', 'go', '450', 'near', 'future', 'b', '...', 'Name',
':', 'message', '...', 'dtype', ':', 'object' ]
[ '3', 'Buy', 'ockpharm', 'as', 'a', 'BTST', '@', '682', 'for', 'a', 'target', 'of', '...', 'Name',
':', 'message', '...', 'dtype', ':', 'object' ]
[ '18', 'buy', 'buy', 'buy...', 'intra', '361', 'Name', ':', 'message', '...', 'dtype', ':', 'object' ]
```

#### 2. Manual labeling

1. By reading the messages which were left and noticing the target that has been posted by the particular user.
2. If the target given by the user is greater than the current price then-BUY and vice versa for SELL.
3. If the target given by the user is near about the current price in range(+ 5or -5) then HOLD.
4. If the message doesn't seem to give any sort of information then NONE.
5. Label Code word used for BUY-0,HOLD-1,NONE-2,SELL-3 as deep learning algorithm can only understand numbers.

## Overview of Data -Sample

user_name	Num_messages	user_level	followers	stocks_tracked	message	Label	Time	ratings
2 vetri	54	Bronze Member	4	84	hold your holdings: it may go 450 near future big announcement on the way		1 about 7 hrs 54 mins ago	0
3 alciprro_2017	105	Silver Member	0	0	Better to exit now and make a reentry below 200 in 2019		3 about 12 hrs 8 mins ago	0
4 chottasid	20	Bronze Member	5	0	★ ★ -- MultiBagger Tips, Commodity Tips, Options Tips, Stock Tips, Futures Tips, For Free. StartNow ★ ★ goo.gl/8		2 about 15 hrs ago	0
5 2525	70	Silver Member	80	51	Buy wockpharm as a BIST @ 682 for a target of 200... Monday market will blast... 📈		2 about 20 hrs 12 mins ago	0
6 csshilarora	5	Bronze Member	1	16	bharti airtel talup with amazone... to lunch low cost mobile... latest news... so dont worry... it will go up		0 about 20 hrs 57 mins ago	0
7 Teet123	1	New Member	0	12	sir this share will go down in upcoming 3 months and will touch near 300		3 about 21 hrs 2 mins ago	0
8 jainpooja	19	Bronze Member	0	10	big fall coming soon... stay away... jio testing is going on for setoff box... after official announcement... big fall!		2 about 21 hrs 25 mins ago	0
9 nishant61	9	Gold Member	39	26	Bharti might see a temporary short covering		2 about 21 hrs 25 mins ago	0
10 2525	70	New Member	0	0	optical fiber advantage of jio won't last ever... rivals are bound to adopt appropriate technologies...		2 about 21 hrs 51 mins ago	0
11 SharmaCapri	1593	Silver Member	5	274	jio wont let it survive...!!		2 about 21 hrs 55 mins ago	0
12 Bulltrail	4	Silver Member	15	12	Only way UP now...shorters beware		2 about 21 hrs 56 mins ago	0
13 Some Stocks	7	Silver Member	35	150	if 354 breaks then it will be down to 320@280/- if Cross 372 then 410-440-445/-		2 about 22 hrs 4 mins ago	0
14 Platinum4all	125	Bronze Member	1	9	please suggest me for i am holding 400 qty @404 so what to do for this... how long i have to wait for profit.		1 about 22 hrs 23 mins ago	0
15 bullishgozp	228	Silver Member	15	12	get set for reverse rise vrrroooooommm		2 about 22 hrs 24 mins ago	0
16 sha_28	484	Platinum Member	30266	0	We are enclosing herewith a press release titled Airtel and Amazon India join hands to introduce a range of afford		2 about 22 hrs 51 mins ago	0
17 harry9581	43	Gold Member	754	0	till now si hit sorry.		2 about 22 hrs 43 mins ago	0
18 bullishgozp	228	New Member	1	3	keep buying for a bright future		0 about 23 hrs ago	0
19 BSE/NSE Announcer	1280284	Gold Member	27650	0	Bharti Airtel has touched a 52-week low of Rs 355.10. At 11:14 hrs, the share was quoting at Rs 355.90, down Rs 10		3 about 23 hrs 4 mins ago	0
20 sk786007	2623	Bronze Member	2	14	buy buy buy... intra 361		0 about 23 hrs 7 mins ago	0
21 p_73439	8	Gold Member	754	0	sell BhartiAirtel future cmp 356 sl 360 tgt 350/347risk reward yours just share my view		3 about 23 hrs 10 mins ago	0
22 abhayverma0105	51	Silver Member	0	0	From Today it will start having new 52 week lows every week. And atleast the next 8 quarters it will report losses w		2 about 23 hrs 16 mins ago	0
23 sk786007	2623	Silver Member	0	0	Sell Reco: Bharti has broken the crucial support of 360 and broken the 52 week low...It is expected to go to 340 in 1		2 about 23 hrs 42 mins ago	0
24 2525	70	Silver Member	4	0	Healthcare Global... HCG buy buy buy!!!! cmp 307, short term target 326... Close to 52week high		0 10:14 AM May 18th	0
25 alciprro_2017	105	Silver Member	166	0	Sell airtel on every rise & sell cmp also 359tgt 320-300		3 10:11 AM May 18th	1
26 alciprro_2017	105	Silver Member	15	12	Anytime U turn possible...		2 9:23 AM May 18th	0
27 less but sure short call	119	Bronze Member	2	8	only buyers in r common same position goes to hear double upper circuit		2 9:19 AM May 18th	0
28 bullishgozp	228	Gold Member	1871	0	Have an opinion on this news? Post your comment here.		2 8:50 AM May 18th	0
29 somnath2018	97	Silver Member	7	0	rcom move 70% in single day. bharti can also but only once it goes to 240 level. guys read my all messages since 4		2 8:32 AM May 18th	0
30 yjyodakhe	38	Silver Member	9	7	want daily 5-6k with 50k investment join my channel by clicking below linkhttps://t.me/joinchat/AAAAA95nlm1kx		2 8:29 AM May 18th	0
31 Market Expert	5503	New Member	1	3	just keep on buying Sunit mittal won't let his company ruin		0 7:01 PM May 17th	0
32 ShainOption	165	Bronze Member	1	0	don't be negative in Bharti Airtel... if Rcom move 65% in a single day than... Bharti Airtel is 100% better than Rcom.s		3 5:29 PM May 17th	0
33 TradeExp	184	Bronze Member	15	0	sell it till 300 levels... only remember that... support levels 361 & 341 again break hajoke to 300 tak sell kro... firm		3 5:27 PM May 17th	1
34 m_73239	1	Bronze Member	4	0	Todays bharti airtel has reached immediate level of 270. There are strong signals present 273. The nearest buy		0 5:32 PM May 17th	0

## Methods

## What exactly Deep Learning is?

-Deep learning is an aspect of artificial intelligence (AI) that is concerned with emulating the learning approach that human beings use to gain certain types of knowledge. At its simplest, deep learning can be thought of as a way to automate predictive analytics.

A type of advanced machine learning algorithm, known as neural networks, underpins most deep learning models. Neural networks come in several different forms, including recurrent neural networks, convolutional neural networks, artificial neural networks and feed forward neural networks, and each has their benefit for specific use cases. However, they all function in somewhat similar ways, by feeding data in and letting the model figure out for itself whether it has made the right interpretation or decision about a given data element.

Neural networks involve a trial-and-error process, so they need massive amounts of data to train on. It's no coincidence that neural networks became popular only after most enterprises embraced big data analytic and accumulated large stores of data. Because the model's first few iterations involve somewhat-educated guesses on the contents of image or parts of speech, the data used during the training stage must be labeled so the model can see if its guess was accurate. This means that, though many enterprises that use big data have large amounts of data, unstructured data is less helpful. Unstructured data can be analyzed by a deep learning model once it has been trained and reaches an acceptable level of accuracy, but deep learning models can't train on unstructured data.

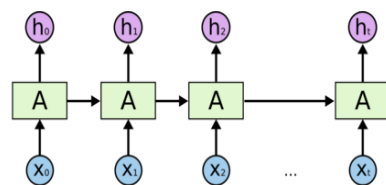
## Limitations of deep learning

The biggest limitation of deep learning models is that they learn through observations. This means they only know what was in the data they trained on. If a user has a small amount of data or it comes from one specific source that is not necessarily representative of the broader functional area, the models will not learn in a way that is generalizable.

Sequence prediction problems have been around for a long time. They are considered as one of the hardest problems to solve in the data science industry. These include a wide range of problems; from predicting sales to finding patterns in stock markets' data, from understanding movie plots to recognizing your way of speech, from language translations to predicting your next word on your phone's keyboard.

With the recent breakthroughs that have been happening in data science, it is found that for almost all of these sequence prediction problems, Long short Term Memory networks, a.k.a LSTMs have been observed as the most effective solution.

LSTMs have an edge over conventional feed-forward neural networks and RNN in many ways. This is because of their property of selectively remembering patterns for long duration of time. The purpose of this article is to explain LSTM and enable you to use it in real life problems.



A typical LSTM network is comprised of different memory blocks called **cells**(the rectangles that we see in the image). There are two states that are being transferred to the next cell; the **cell state** and the **hidden state**. The memory blocks are responsible for remembering things and manipulations to this memory is done through three major mechanisms, called **gates**.

## Structure Of LSTM Model Created For This Project

- Tokenizer API
- Encoding with one\_hot
- Sequence Padding
- SMOTE
- Embedding Layer
- Activation Function('Softmax')
- Loss Function( 'Categorical\_crossentropy')
- Optimizer Function('RMSprop')

## EXPERIMENTS AND RESULTS

<b>Language</b>	<b>Python 3.5</b>
<b>Framework</b>	<b>Keras</b>
<b>Windows</b>	<b>10</b>
<b>Software</b>	<b>Jupyter Notebook</b>

Types of Data	Activation function and loss	Optimizer	Epochs	Batch_size	F1_score	Precision	Recall
Airtel(replicated )	Softmax& categorical	RMS_prop	75	20	0.82	0.84	0.85
Airtel(SMOTE)	Softmax& categorical	RMS_prop	50	16	0.66	0.66	0.66
Airtel(without SMOTE)	Softmax& categorical	adam	51	20	0.70	0.68	0.73
Airtel(replicate with k-fold=10)	Softmax& categorical	RMS_prop	50	20	0.95	0.97	0.95

Fig.1 Airtel

Types of Data	Activation function and loss	Optimizer	Epochs	Batch_size	F1_score	Precision	Recall
Tata_steel(replicated)	Softmax&categorical	RMS_prop	75	20	0.82	0.83	0.85
Tata_steel(SMOTE)	Softmax&categorical	RMS_prop	50	16	0.49	0.51	0.51
Tata_steel(without SMOTE)	Softmax&categorical	adam	75	20	0.69	0.65	0.70
Tata_steel(replicate with k-fold=10)	Softmax&categorical	RMS_prop	50	20	0.93	0.95	0.96

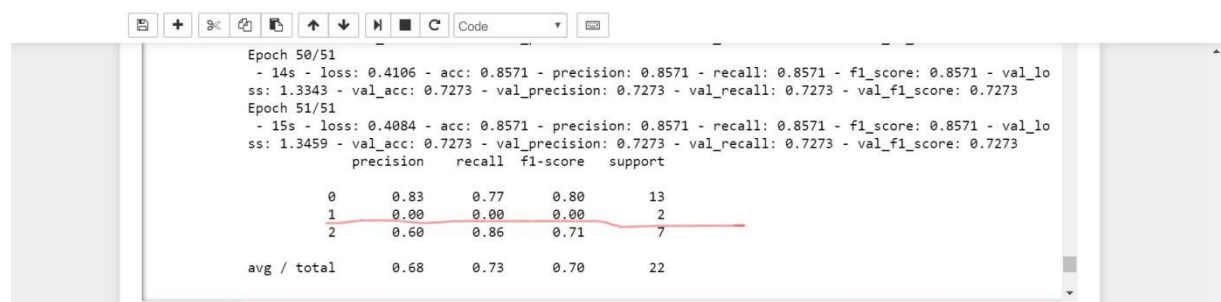
Fig.2 Tata\_steel

Fig.3 Pc\_Jewellers

Types of Data	Activation function and loss	Optimizer	Epochs	Batch_size	F1_score	Precision	Recall
Pc_Jewellers(replicated)	Softmax&categorical	RMS_prop	75	20	0.79	0.80	0.82
Pc_Jewellers(SMOTE)	Softmax&categorical	RMS_prop	50	16	0.54	0.59	0.58
Pc_Jeweller(without SMOTE)	Softmax&categorical	adam	51	20	0.61	0.60	0.60
Pc_Jewellers(replicate with k-fold=10)	Softmax&categorical	RMS_prop	50	20	0.90	0.91	0.91

### STEPS INVOLVED IN EXPERIMENTS:-

1.As we started the experiment with Airtel Data set we took the data set without applying smote and k-fold and using 'adam' as optimizer and 'softmax' and 'categorical\_crossentropy' as activation function and loss function respectively. The result that we found is shown in table 1. But we faced a problem i.e the class 'HOLD'(HOLD-1) didn't show any result. As you can see down here in the image class 1 holds no calculation for precision ,recall and f1-score.



```

Epoch 50/51
- 14s - loss: 0.4106 - acc: 0.8571 - precision: 0.8571 - recall: 0.8571 - f1_score: 0.8571 - val_loss: 1.3343 - val_acc: 0.7273 - val_precision: 0.7273 - val_recall: 0.7273 - val_f1_score: 0.7273
Epoch 51/51
- 15s - loss: 0.4084 - acc: 0.8571 - precision: 0.8571 - recall: 0.8571 - f1_score: 0.8571 - val_loss: 1.3459 - val_acc: 0.7273 - val_precision: 0.7273 - val_recall: 0.7273 - val_f1_score: 0.7273
precision    recall  f1-score   support

 0         0.83         0.77         0.80         13
 1         0.00         0.00         0.00          2
 2         0.60         0.86         0.71          7

avg / total         0.68         0.73         0.70         22

```

2. So next step that we took to solve this problem was we applied k-fold in our fold but the same thing happened with the HOLD class though there other class showed some better result but we didn't get our desired result.

3. In next step we applied SMOTE to the data set and by the help of this method we actually got better result i.e the class HOLD had some calculation.

```
3. 1.3727 - val_acc: 0.5540 - val_precision: 0.6004 - val_recall: 0.5004 - val_f1_score: 0.5555
Epoch 48/50
- 4s - loss: 0.1981 - acc: 0.9466 - precision: 0.9525 - recall: 0.9379 - f1_score: 0.9450 - val_loss: 2.1644 - val_acc: 0.5517 - val_precision: 0.5640 - val_recall: 0.5345 - val_f1_score: 0.5484
Epoch 49/50
- 4s - loss: 0.1732 - acc: 0.9483 - precision: 0.9597 - recall: 0.9345 - f1_score: 0.9465 - val_loss: 1.8141 - val_acc: 0.6379 - val_precision: 0.6401 - val_recall: 0.5862 - val_f1_score: 0.6111
Epoch 50/50
- 4s - loss: 0.1573 - acc: 0.9586 - precision: 0.9647 - recall: 0.9448 - f1_score: 0.9540 - val_loss: 1.6670 - val_acc: 0.6552 - val_precision: 0.6594 - val_recall: 0.6293 - val_f1_score: 0.6437
precision    recall    f1-score   support
0           0.61         0.66         0.63         29
1           0.65         0.69         0.67         29
2           0.61         0.59         0.60         29
3           0.77         0.69         0.73         29
avg / total         0.66         0.66         0.66        116
```

## CONCLUSION

- In this project, we have built an algorithm for prediction of stock movements using LSTM model.
- We have calculated performance metrics of the system using this model.
- The experiment results show that precision, recall and f1-score values for SMOTE data with 'Softmax' activation function and 'Categorical' cross entropy are 66%, 66% and 66% respectively.
- Data without SMOTE 'Softmax' activation function and 'Categorical' cross entropy are 68%, 73% and 70% respectively. Our next step is to further increase the performance metrics of the system.

