Modelling and Prediction of Indian Stock Market using relevant news data and other covariates.

Alimpan Barik and Nilabjanayan Bera

Ramakrishna Mission Vivekananda Educational and Research Institute, Belur Math,Howrah.

`Finance Project`

October 30, 2021

This project aims to predict the stock price using relevant news data and some other covariates. Usually ARCH-GARCH model is implemented to predict stock prices from the past - which are conventional Time Series approaches. Since, stock market data is volatile in nature,so we can't model perfectly using the conventional time series approach. Deep learning methods are proved to be useful in different types of forecasting tasks. Here we want to propose a method inspired from Deep learning. Also , news and social media have a huge impact on the market. We want to study them as well and build a model which would include the effect of news and social media trends in the stock price.

1 Data Collection

We have to collect the data manually from the internet as we do not intend to do our project on some ready-made dataset. We propose to have data of approx. 2.5 years-2 years for training and 6 months for testing. There are 3 steps of data collection-

- **Stock Price Data collection:** We have collected the Indian stock price data from yahoofinance.com
- **News Data collection:** We have used web scraping techniques in python to extract news data from various websites.
- **Tweet Data collection:** Tweet data is collected from twitter (using the tweepy package)

2 Data Preprocessing

For the pre-processing part we have to tackle 3 different data in 3 different ways.

- **Stock Data** We have to normalize the stock data to fit it in an LSTM model.
- **News Data** To calculate the sentiment scores of the news data we propose two different ways.First, using the BERT model we encode the news and get some relevant score (positive or negative according to its sentiment) and alternately, using some pre-trained model (on some available news sentiment dataset) we get the sentiment scores.
- **Twitter Data:** Similarly, we classify the tweets in three different classes according to its sentiment-(+1,0 and -1)

3 Training:

We will use the LSTM model for training- where will have the input vectors and the model will give us an output . The input vector consists of past stock prices of few days, and news score ,tweet class as covariates. We plan to have a pilot run based on a small dataset (maybe for a few months) and set our plans accordingly within 06.11.2021.

4 Testing:

We shall evaluate the performance of our model on some new data (maybe for one month) for our pilot run.We also propose to compare our model with the ARCH-GARCH model and the standard LSTM model (without the effect of news and twitter data) so that we can prominently understand the effect of news and social media on the stock market- which is our primary objective. .