Stock Market prediction and analysis using various algorithms

By Satya O. Aakash and R. Uma Sai Charan

Abstract:

Stock market is one of the key factors of the economy. It is always a n important topic for researchers from both technical and financial areas. It is one of the most challenging areas for investors as there are a number of factors which affect the stocks directly and indirectly. It is impossible to predict the stock market with 100% accuracy. We analysed various statistical and machine learning algorithms like ANN, regression algorithms SVM, LSTM and sentimental analysis and this is a detailed review of those algorithms.

1.Introduction:

Stock market is the aggregation of buyers and sellers of stocks which represent the ownership claims on business. In a country like India where there are thousands of legally recognized companies, the stock market plays a major role in the country's economy. Due to the large number of factors affecting the stock market especially in today's world where even a simple tweet can affect the rates of stocks. This makes it very difficult for stock market investors and stock brokers. As this is a very hot topic always as this includes cores of business a lot of research was done especially in the last decade. It is always a n important topic for researchers from both technical and financial areas. It is impossible to predict the stock market with 100% accuracy. Our main motive is to analyse and give a detailed report of most popular algorithms.

2.Literature survey:

Michitaka Kosaka et al.[1] proposed a neural network model for the technical analysis of stock market Prediction for Tokyo Stock Price Index (TOPIX). They worked on increasing the accuracy of the prediction. In a way they improvised the existing algorithm by adding some learning algorithms and one can actually calculate the rate of profit in buying conditions using this equation and the rate of profit in selling situations using.

ASHISH SHARMA et al.[2] consists of a detailed analysis of stock market prediction using different types of machine learning regression algorithms. Various types of regression like Polynomial regression , RBF regression and sigmoid regression and linear regression . linear regression can be used to fit a predictive model to an observed data set of the two variables.

Hiransha et al. [3] used deep learning algorithms for predicting the stocks of NSE and NYSE by training four different algorithms MLP, RNN, LSTM, CNN with the stock price of TATA motors from NSE and various other companies. They proved the same algorithms can be used by different organizations like NSE and NYSE which proves that even though they are different they undergo the same basic dynamics. Also apparently CNN gave better and more accurate results than the other three algorithms and the main reason is capable of capturing the abrupt changes in the system since a particular window is used for predicting the next instant.

Ankit Thakkar et al.[4] proposed that nonlinear predictions of stock marketing have also been very effective especially in the last decade. They propose that it is necessary to elaborate the necessity of applying fusion in the stock market. They used three major kinds of fusions: information fusion, feature fusion, and model fusion. The target area where fusion is integrated can decide on its impact on the overall prediction performance as well classification-based categorization and characteristics of the articles may be challenging. So we need other potential fusion techniques that have been applied to the financial market, followed by the potential future directions based on fusion in the stock market.

Kelotra*,et al.[5] subjected the data obtained from the live stock market to the feature extraction process from which the required features needed in the prediction of the stock market are extracted. The effective features are fed to the Deep-Conv LSTM model that is trained by using the proposed Rider-MBO algorithm, which is developed from the integration of the ROA and the MBO.

Xiongwen Pang et al.[6] developed an innovative neural network approach to achieve better stock market predictions. The LSTM neural network with embedded layer and the long short-term memory neural network with automatic encoder are proposed on the basis of LSTM neural network.

Jonathan L.Ticknor et al.[7] proposed that it will be effective if the model contains Bayesian regularization with Leven –Berg –Marquardt algorithm to forecast the movement of socks with minimum error margin.

Aditya Gupta et al.[8] present a HMM based MAP estimator for stock prediction. This model uses a latency of some days to predict the stock value for the *(d + 1)st* day. A MAP decision is made over all the possible values of the stock using a previously trained continuous-HMM. We assume four underlying hidden states which emit the visible observations (fractional change, Fractional high, fractional low).

Shen et al. [9] proposed that the use of data collected from financial markets with various algorithms to predict the stock movements and his trained model declares that correlation analysis shows a strong connection between the US stock market and global markets. The connection between US stock and global markets that close before or after US trading time. They proposed various algorithms in the context of machine learning to predict daily based US stocks and also the results show high accuracy. The proposed model generates higher profits than daily benchmarks.

Pathak et al. [10] proposed that the existing work may be developed into a more accurate model to predict the stock market accurately. The model developed by defining refined fuzzy rules, improving training data, and also time frame shows results in better prediction.

The proposed plan indicates the returns or investments in real-time. The model includes the technical aspects like historical data with machine learning, news headlines with sentimental analysis then combining with and operator to the fuzzy logic module and they believed that this shows the best accuracy for returns or investments in real-time.

Vishwakarma et al. [11] proposed a model called Arima in the time series module to use the past data to understand its pattern and also they proposed another model called Sarima which tends to capture the trend in time series and forecast the values for future dates. They also proposed a model named Prophet which is also used to capture the trend and according to seasonalities they try to pretend the future values of the stock market. They failed to find good accuracy with these models and they declared that the stock market mainly depends on upcoming markets and how the stocks own them a price too high or low and also they declared that their own models are not good for forecasting by using Arima, Sarima, and prophet for the stock market. Finally, they declared by end of their research stock market cannot be predicted through time series forecasting and past data.

Alvarez et al. [12] worked on artificial intelligence techniques and machine learning algorithms. They proposed an application called WEKA which contains different algorithms Which could

provide many economical insights related content to society. This tool helps with sentimental analysis i.e the tool analyses how the market was going and how the stocks were exchanging in the market, they analyse the people tweets and uses all these insights to predict the future stocks with more accuracy and to provide to society in a single tool.

Parmar et al. [13] have contributed an LSTM Model to determine the future stock prices and by implementing machine learning techniques with the LSTM model and proved that the model yielding positive results in accuracy and also they can see better efficiency in the model. By this, they declared that the results concluded that it is possible to predict the stock market with more accuracy and efficient models using machine learning techniques.

Reddy, V.K.S [16] proposed an algorithm using data collected from financial markets and predicted stock index movements. In this paper, he took SVM as an opponent to his proposed model and predicted with higher efficiency than SVM. This model also generates higher profits than selected models.

3.Applications:

3.1 Support Vector Machine

A Support Vector Machine (are also called support-vector networks) implements the classification by finding the hyperplane that maximizes the margin between two classes [14].

Here the hyperplane represents the support vectors. Support Vector Machine is the best and suitable algorithm for time series prediction. The support vector machine proves that it is one of the best algorithms by a high accuracy model with some good feature selection from data. The algorithm shows more accuracy with good feature selection than selected benchmark models (i.e fixed algorithms).

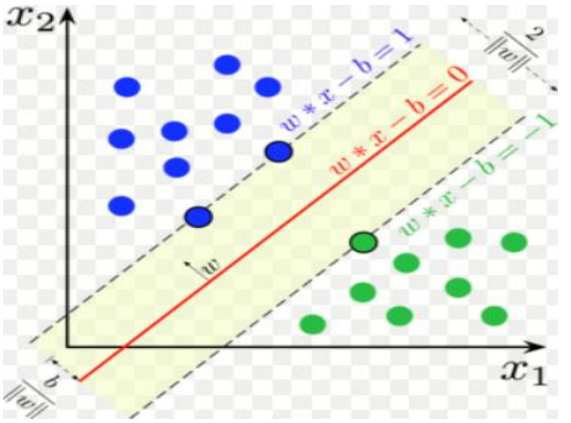

Fig.1 explains SVM in Stock Market Prediction [source: Wikipedia]

From Fig.1 the blue points (we also called as positive hyperplane) can be referred to as one type of likely data from the stock market and the green points (we also called as negative hyperplane) can be referred to as another likelihood data from the stock market, here the classification of support vector machine is done by hyperplane which is that red line. The equations and the magnitude values represent the hyperplane and the classification classes.

3.2 LSTM (Long Short Term Memory) Network-Based Model

Long Short-term memory is an artificial recurrent neural network architecture, used in the field of deep learning. LSTM has advanced features than other algorithms that cannot only process single data points (images), but also entire data (like speech or video). LSTM is used for handwriting recognition, speech recognition, anomaly detection in network traffic, etc.

An LSTM consists of a cell, an input gate, an output gate, and a forget gate. the method of LSTM describes how the cell remembers values over time intervals and therefore the three gates (input, output, forget gate) regulate the flow of data in and out of the cell [18]. LSTM is well known for the best classification and processing algorithms used and also LSTM is best in time series predictions like a stock market prediction. There can be lags of unknown duration which can be encountered during traditional RNNs. LSTM had the advantage of a memory unit when compared to a common recurrent unit.

The main purpose to use this model in stock market prediction is because it depends on a large amount of data and also it depends on long history which is to be stored and also to retain the information that gives the edge by LSTM over Traditional RNNs. LSTM proves in the past that it predicts with more accuracy and also more reliability when compared to other methods.

LSTM prevents the problem of Vanishing Gradient which occurs because the stock market involves in the processing of huge data so the gradients with respect to the weight matrix may become very small and may degrade the learning rate.[13]

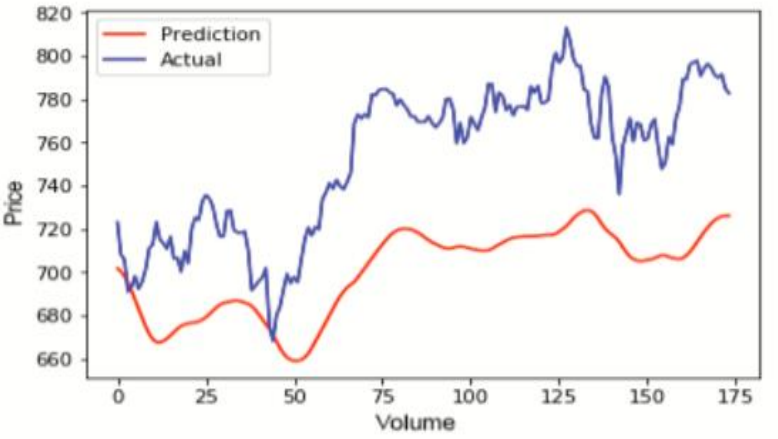

Fig.2 [13] shows the graph between predicted and actual stock market analysis in price and volume.

3.3Sentimental analysis module

Sentimental analysis is a text classification tool which analyses that the incoming message or tweet of a person and tells the statement is positive, negative, or neutral. The module works in a flow that we input a message and it responds in a way it is positive, negative, or neutral and the sentiment will be stored as data and it is used for a training model and predictions of the stock market.

Sentimental analysis is a material and also help a business to understand the social sentiment of their product or service. In these terms, the stock market prediction results with more accuracy[10]. In recent technologies, this module will play a bigger role along with machine learning and deep learning algorithms to predict the stock market which mainly depends on brands and business investments.

The sentimental analysis contains the work to be done in the following ways

1.DataCollection	Collect data from trusted sites or dataset

2.Tokenizing	Each news headline or social media message, or tweets will be broken down into sentences then into words
3.Lemmatizing	Process of reducing the inflected word their root word
4.Finding the most informative features	Store the word which contributes the most positivity or negativity word to a sentence
5.classifying features into positive and negative	Classified into positive and negative and transferred to respective packages
6.Adding these features to the sentiment analyzer lexicon	Words are then added to the sentiment analyser with the strength of positive or negative
7.classifying the testing data into positive and negative sentiments using the training set	Classification completed in this step and the data will be ready with the classification of positive or negative or neutral.

Table 1 table for the sentiment analysis process [10]

Fig 3 sentimental analysis

3.4 Artificial Neural Networks:

ANN is a computational structure which performs in a similar manner to that of biological neurons.[3] It is considered as an efficient optimization tool for predicting the time series, and also, it predicts the hidden and the unknown records.[5] The most important advantage of ANN is its capability to learn the underlying patterns from the data, where most of the conventional methods fail [3]. A commonly used performance criterion function is the sum of squares error function [9]. Where, p represents the patterns in the training set, yp is the output vector (based on the hidden layer output), tp is the training target.[9]

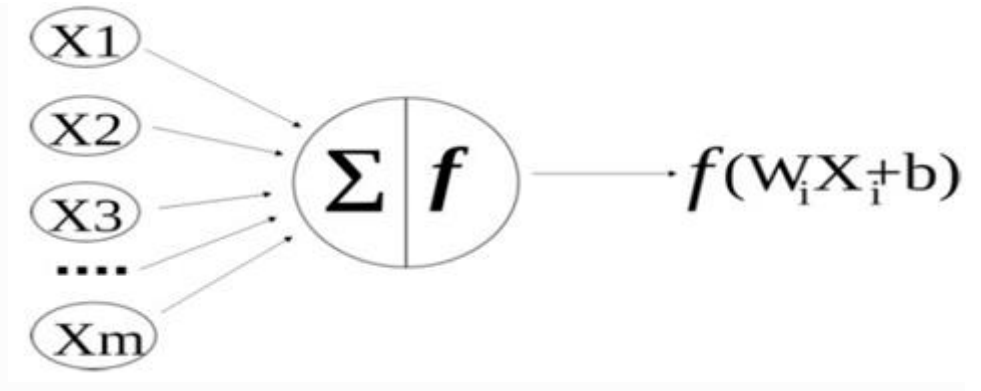

Fig1: Artificial neural cell

ANN includes a set of threshold functions. These functions trained on historical data after connecting each other with adaptive weights and they are used to make future predictions.

Fig-1 is the rough sketch of an artificial neuron and the x1 x2…..xn are the inputs which we give and the sigma function and the other function train themselves with the given data and predict an output. The neuron adds the inputs and multiplies them with their weights using the following equation :$A= \sum x_i w_i + b$ [6]

3.5 Multi layered perceptron:

Multi layered perceptron also called as Feed Forward Network [3] are the major part of a neural network. In MLP each input node is connected to a hidden layer neuron through a weighted matrix w (i) [3]

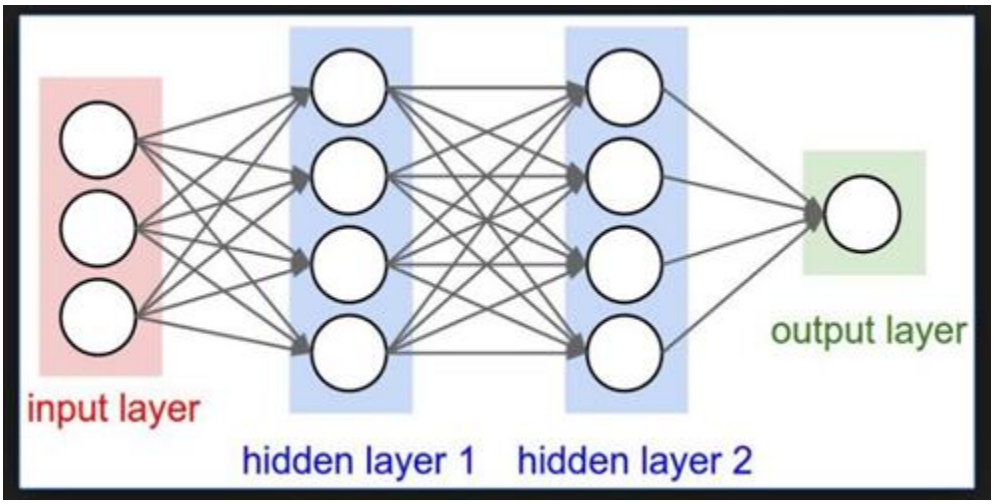

Fig-2 (Rough sketch of how an ANN works)

Each of these nodes receives inputs from previous node and artificial neuron is the one present in hidden and output layers. . Fig-2 explains how the MLP works, we give our data as the input in the input layer and the neuron trains itself in the hidden layer-1, 2 and predicts an output through the output layer. The equation of the activation function of an I th equation is

$$h_i = f(\mu_i) = f(\sum_{k=0}^{k} w_{ki} x_k)$$

Where

h_i =ith hidden neuron

$f(\mu_i)$ = link function which provides non-linearity between input and hidden layer

w_{ki} = weight in the (*k, i*)*th* entry in a (K X N) weight matrix

x_k : *Kth* input value

3.7 Regression:

Regression analysis is a predictive modelling technique that analyses the relation between the target or dependent variable and independent variable in a dataset. Depending on weather the variable is linear or a non-linear there are different types of regression algorithms we can use.

Regression analysis helps us to understand how the typical value of the in-dependent variable affects the value of any one dependent variable when it changes while the other independent variables are held fixed. Regression analysis estimates the conditional expectation of the dependent variable given the independent variables.[2]

Types of regressions:

1) Polynomial regression

2) Linear Regression

3) RBF Regression

3.8 Polynomial Regression:

Polynomial regression is a type of regression in ML which is similar to Multiple Linear Regression with some key modifications. In this regression the relationship between independent and dependent variables, that is X and Y, is denoted by the n-th degree. It fits a nonlinear relationship between the value of x and the corresponding conditional mean of y denoted E(y | x) Least mean squared method is used in polynomial regression. The aim of regression analysis is to model the expected value of a dependent variable y in terms of the value of an independent variable (or vector of independent variables) x.[19]

Fig-4(predictive analysis using polynomial regression)

The following equation is used $y= a_0+a_1x+ \varepsilon$, which is the general form of 1-degree polynomial equation with an unobserved error **ε.**[6] In case in any condition if the above equations does not hold we can increase the degree of the polynomial and add the error margin and we can use this equation $y= a_0+a_1x+a_2x^2+\varepsilon$ [2] In general we can derive the n-th degree polynomial of y as follows $y= a_0+a_1x+a_2x^2+a_3x^3+\ldots.+a_nx^n+\varepsilon.$

Linear Regression:

linear regression is a linear approach to modelling the relationship between a dependent variable and one or more explanatory variables (or independent variables).[19]

The linear regression model consists of an independent variable and a dependent variable related to each other in a linear way. The linear regression is denoted by the equation y=mx+c+e

Where

m= slope of the graph

c= intercept

e=error margin

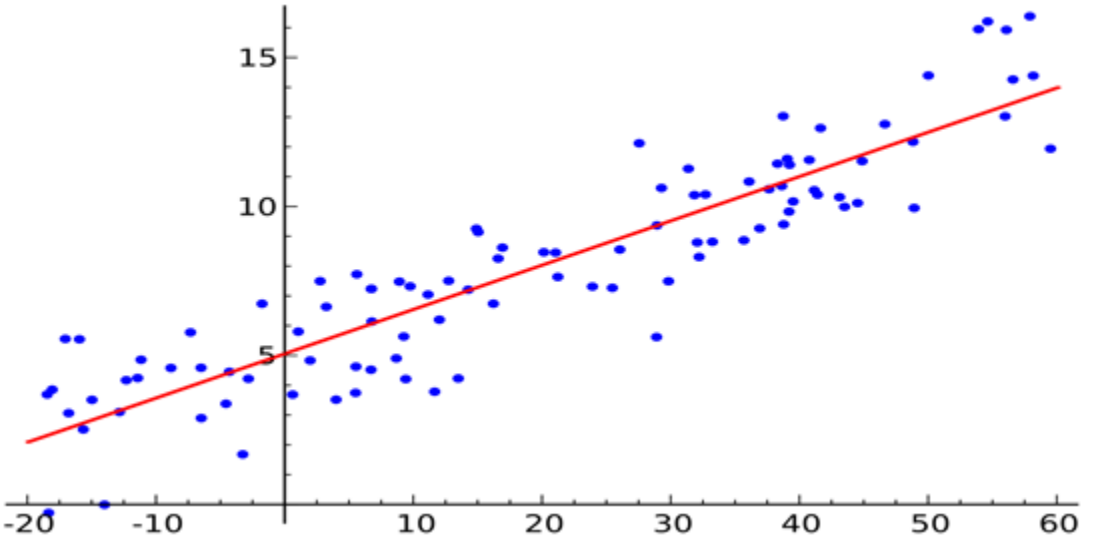

Fig-5: sample graph of linear regression

RBF regression:

A radial basis function (RBF) is a real-valued function. Its value largely depends on its distance from the origin .so that $\Phi(x) = \Phi(||x||)$ It is depended on either origin or some other point 'c' called centre, so that $\Phi(x,c)= \Phi(||x-c||)$ this is why norm is also known as Euclidean Distance, although other distance functions are also possible. For instance, when Łukaszyk–Karmowski metric is used it it most likely for some radial functions so that we can avoid problems with bad conditions of the matrix to solve to determine coefficients [2].An RBF Radial Basis Function net is similar to a 2-layer network. In this function the input is completely connected to a hidden layer. Then, we take the output of the hidden layer and perform a weighted sum to get our output.

4. Languages, Tools, and libraries

Python, R, and Java are used for machine learning algorithms for stock market prediction. Each language has its own features and also the statistics in a survey conducted by stack overflow declared python is the most popular and fastest-growing language in 2020.[14]

Sci-kit learns can be used for data mining and analysis. Pandas is used for labelling data and NLTK is used for sentimental analysis.[14]
Matplotlib is used as a visualization tool for labelled data in pandas and plotly works as an online analytics and data visualization tool. [14]

Conclusion:

This paper summarizes important algorithms in machine learning, deep learning which are used for stock market prediction, and also the paper discusses how the techniques are changed from day to day in stock market prediction. This paper discusses what languages and libraries can be used for stock market prediction and also recommends the algorithms discussed in this paper for more accurate results. In the future, the prediction of the stock market will emerge with new technologies like a combination of algorithms with some good feature selection in data sets to give more accurate results and also for processing huge data.

References:

1) Mizuno, H., Kosaka, M., Yajima, H., & Komoda, N. (2001). Application Of Neural Network To Technical Analysis Of Stock Market Prediction.
Application Of Neural Network To Technical Analysis Of Stock Market Prediction

2) A. Sharma, D. Bhuriya and U. Singh, "Survey of stock market prediction using machine learning approach," 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, 2017, pp. 506-509, doi: 10.1109/ICECA.2017.8212715.

3) Hiransha, M., E. Ab Gopalakrishnan, Vijay Krishna Menon, and K. P. Soman. "NSE stock market prediction using deep-learning models." *Procedia computer science* 132 (2018): 1351-1362.

4) Thakkar, A., & Chaudhari, K. (2020). Fusion in stock market prediction: A decade survey on the necessity, recent developments, and potential future directions. *Information Fusion*, *65*, 95-107.

5) Kelotra, Amit, and Prateek Pandey. "Stock market prediction using optimized deep-convlstm model." *Big Data* 8, no. 1 (2020): 5-24.

6) Pang, Xiongwen, Yanqiang Zhou, Pan Wang, Weiwei Lin, and Victor Chang. "An innovative neural network approach for stock market prediction." *The Journal of Supercomputing* 76, no. 3 (2020): 2098-2118.

7) Ticknor, Jonathan L. "A Bayesian regularized artificial neural network for stock market forecasting." *Expert Systems with Applications* 40, no. 14 (2013): 5501-5506.

8) Gupta, Aditya, and Bhuwan Dhingra. "Stock market prediction using hidden markov models." In *2012 Students Conference on Engineering and Systems*, pp. 1-4. IEEE, 2012.

9.Shen, S., Jiang, H., & Zhang, T. (2012). Stock market forecasting using machine learning algorithms. *Department of Electrical Engineering, Stanford University, Stanford, CA*, 1-5.

10.Pathak, A., & Shetty, N. P. (2019). Indian stock market prediction using machine learning and sentiment analysis. In *Computational Intelligence in Data Mining* (pp. 595-603). Springer, Singapore.

11.Vishwakarma, A., Singh, A., Mahadik, A., & Pradhan, R. Stock Price Prediction Using Sarima and Prophet Machine Learning Model.

12.Hernández-Álvarez, M., Hernández, E. A. T., & Yoo, S. G. (2019, February). Stock Market Data Prediction Using Machine Learning Techniques. In *International Conference on Information Technology & Systems* (pp. 539-547). Springer, Cham.

13.Parmar *et al*., "Stock Market Prediction Using Machine Learning," *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)*, Jalandhar, India, 2018, pp. 574-576, doi: 10.1109/ICSCCC.2018.8703332.

14.Pahwa, N., Khalfay, N., Soni, V., & Vora, D. (2017). Stock prediction using machine learning a review paper. *International Journal of Computer Applications*, *163*(5), 36-43.

15.Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock market index using the fusion of machine learning techniques. *Expert Systems with Applications*, *42*(4), 2162-2172.

16.Reddy, V. K. S. (2018). Stock market prediction using machine learning. *International Research Journal of Engineering and Technology*, *5*(10).

17) Vaisla, Kunwar Singh, and Ashutosh Kumar Bhatt. "An analysis of the performance of artificial neural network technique for stock market forecasting." *International Journal on Computer Science and Engineering* 2, no. 6 (2010): 2104-2109.

18) https://en.wikipedia.org/wiki/Long_short-term_memory

19) Wikipedia