International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 11 | Nov 2019 www.irjet.net p-ISSN: 2395-0072

# Enhancement in Financial Time Series Prediction With Feature Extraction in Text Mining Techniques

**Assistant Prof. Anjali Sanjivanrao More[1], Deepa Sunil Ranaware[2], Bhakti Dattatraya Wamane[3], Gouri Shivaji Salunkhe[4]**

*[1]Assistant Professor of Computer Engineering &SavitribaiPhule Pune University*
*[2,3,4]Pursuing Bachelor of Computer Engineering &SavitribaiPhule Pune University*
*Suman Ramesh Tulsiani Technical Campus Faculty of Engineering Kamshet, Pune, India*

---

**Abstract -** *News has been a very important supply for several monetary statistic predictions supported elementary analysis. However, digesting an enormous quantity of reports and information revealed on the net to predict a market will be heavy. This paper introduces a subject model supported latent Dirichlet allocation (LDA) to get options from a mix of text, particularly news articles and monetary statistic, denoted as monetary LDA (FinLDA). The options from FinLDA are served as extra input options for any machine learning algorithmic program to boost the prediction of the monetary statistic. The Proposed System offer posterior distributions employed in chemist.*

**Key words**-: **Bayesian method, Data Mining, Latent Dirichlet Algorithm(LDA), Stock market, Support Vector Machine(SVM).**

**Terminologies-**:

LDA – Latent Dirichlet Algorithm SVM – Support Vector Machine

FinLDA – Financial Latent Dirichlet Algorithm

EMH - Efficient market hypothesis RSI - Relative Strength Index

A/DG-Accumulation/ Distribution generator

## 1. INTRODUCTION

Efficient market hypothesis (EMH) developed by Fama[1][2] advised that worth changes instantly answer new info, and that they square measure unpredictable. consequently, historical knowledge can't be accustomed build profitable predictions. However, several approaches are accustomed predict monetary market movement, crashes or booms[3], and also the prediction still remains the topic of active continued analysis. Basically, technical and basic analyses square measure utilized by investors to predict monetary time evolution, like stock costs. Technical analysts believe that historical market knowledge, primarily worth and volume, offer options for worth prediction[4]. Also, worth and volume may be extended to a lot of complicated indices, like relative strength index (RSI), Accumulation/ Distribution generator (A/DG), etc. Technical analysis focuses on victimization ways to extract different info from the historical worth and volume. In distinction, varied knowledge sources may be employed in basic analysis; they'll be any info a few company or its sector, e.g., income magnitude relation, come back on assets (ROA), etc., or economics, e.g., America gross national product, America shopper indicant (CPI), etc.

Moreover, the elemental knowledge may be unstructured matter knowledge, e.g., international news articles, messages in a very net board, public company disclosures, etc., from that square measure tougher to extract info. consequently, monetary models for stock prediction square measure typically supported numerical technical and basic knowledge and specialise in modeling to boost the results, e.g., ARCH models, GARCH models, machine learning algorithms, etc. [5]– [8]However, once text mining had emerged and become sensible to extract info from text, monetary analysis took the unstructured matter info under consideration a lot of typically. several relied on recognizing keywords: Wüthrich et al. [9], as an example, extracted articles printed on The Wall Street Journal web site supported lists of keywords records, judged to be important by domain specialists. Then, the keyword counts were weighted and utilized by their rules, applied to predict the 1-day trend for 5 major equity indices. Although taking news into thought, some studies set matter info aside and solely used its numerical knowledge, e.g., the amount of stories articles and their timestamps. Chan [10], as an example, used the date of the news on that the stock was mentioned (stock name was used as a keyword) and located proof of `post-news drift'. Their knowledge supported activity finance theory [11], [12]regarding each capitalist over and under-reaction to new info returning from capitalist unreason. Some studies thought-about each word, instead of just some keywords. for example, Fung et al. [13]

investigated the immediate impact of stories articles, archived by Reuters, on the value changes of city exchange (SEHK) stocks and bestowed a system to predict rise and fall trends of stock price. They delineate radio- controlled bunch to strain articles that weren't helpful in trend prediction. each word within the article was extracted associate degreed allotted a numerical price by tf-idf[14] so a number of the articles were filtered out by an extension of progressive k- Means bunch. Then, they were distinguished by a brand new differentiated coefficient theme to become options in a very Support Vector Machine (SVM) to predict the trends.

## 2. Literature Review

Amirhessam Tahmassebi et al.(2018),"iDeepLe: Deep Learning in Deep Learning in a Flash",may 2018,IEEE

a. **Methodology**: iDeepLe was written in Python with the help of various API libraries such as Keras, TensorFlow, and Scikit-Learn.

b. **Findings & Application :** A powerful deep learning pipeline, deep learning (iDeepLe) was proposed for both regression and classification tasks.

c. **Remark( future Scope & Conclusion):** In Future deep learning industry will adopt a core set of standard tool[1].

**[2] Stefan Feuerriege, Antal Ratku, Dirk Neumann et al. (2016), "Analysis of how underlying topics in financial news affect stock prices using latent Dirichlet allocation", 2016, IEEE[2]**

a. **Methodology**: This method was used the text mining process and with the help of this Noise was removed.

b. **Findings and Application:** Identified the factors that influence stock price changes has always been a critical research question.

c. **Remark (Future scope and conclusion):** In future content of news announcements convey information that will processed and subsequently reflected in stock market prices.

**3. Amir H. Gandomi, Anke Meyer- Baese et al.(2018), "A Pareto Front Based Evolutionary Model for Airfoil Self-Noise Prediction",2018,IEEE :**

a. **Methodology:** A GP model based on multi-objective genetic programming approach and NSGA-II was applied.

b. **Findings & Application :** Total five features in the database were used as the inputs of the GP model for the symbolic regression task.

c. **Remark( future Scope & Conclusion):** In Future developed an evolutionary symbolic implementation for airfoil self-noise prediction can be developed[3].

**Y. Guo, S. Han, C. Shen, Y. Li, X. Yin, and Y. Bai et al.(2018), "An Adaptive SVR for High-Frequency Stock Price Forecasting", March 2018,IEEE:**

**a. Methodology :** With the development of information technology, the research on finance changed from macro to micro.

**b. Findings & Application :** SVR was proposed for stock data at three different time scales, including daily data.

**c. Remark( future Scope & Conclusion) :** In future an adaptive SVR based on PSO is proposed to enhanced the versatility of the model and to avoid suffering from adjusting parameters of SVR[4].

**[5] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin et al.(2019)," CatBoost: unbiased boosting with categorical features", Jan 2019,IEEE:**

a. **Methodology :** Represented the key algorithmic techniques behind CatBoost, a new gradient boosting toolkit.

b. **Findings and Application**: This method was used the text mining process and with the help of this Noise was removed.

c. **Remark (Future scope and conclusion):** In future content of news announcements convey information that will processed and subsequently reflected in stock market prices.

## 3. EXISTING SYSTEM

In existing studies set matter data aside and solely used its numerical knowledge, e.g., the quantity of stories articles and their timestamps. as an example, used the date of the news on that the stock was mentioned (stock name was used as a keyword) and located proof of `post-news drift'. Their knowledge supported activity finance theory concerning each capitalist over and under- reaction to new data coming back from capitalist insanity. Some studies thought of each word, instead of just some keywords. for example, Fung et al. investigated the immediate impact of stories articles, archived by Reuters, on value changes of Hong Kong securities market (SEHK) stocks and bestowed a system to predict rise and fall trends of stock price. They

delineate target- hunting bunch to strain articles that weren't helpful in trend prediction. each word within the article was extracted Associate in Nursingd appointed a numerical price by tf- idf then a number of the articles were filtered out by an extension of progressive k-Means bunch. Then, they were distinguished by a brand new differentiated coefficient theme to become options during a Support Vector Machine (SVM) to predict the trends.

**DISADVANTAGES**

1. Only used its numerical data.

2. 2. It considered every word, rather than only some keywords.

## 4. PROPOSED SYSTEM

The Proposed System deals with a replacement domain-specific topic model, the FinLDA model. It deals with incorporating changes in money statistic into the common Latent Dirichlet Allocation to come up with a replacement set of latent topics associated with the changes in an exceedingly statistic. Here The Proposed System have a tendency to delineate 2 variant of FinLDA: 1) distinct FinLDA (d- FinLDA) uses, as input, the movements that square measure changes classified into a distinct set of values (e.g., no change, considerably up, minor up, etc.), whereas 2) continuous FinLDA(c-FinLDA) uses real numbers or actual variations. The Proposed System have a tendency to provided posterior distributions utilized in chemist sampling for parameter estimation and logical thinking in topic modeling with FinLDA. The Proposed System have a tendency to thought of FinLDA to be a feature extraction in data processing. As a result, this text focuses on the feature extraction in an exceedingly information preparation part, however The Proposed System have a tendency to still would like the opposite phases in data processing to urge the ultimate results. consequently, The Proposed System have a tendency to provided the small print of a framework for applying FinLDA in text and data processing for money statistic prediction. The Proposed System have a tendency to used 2 approaches to organize datasets for analysis, i.e., train-test split and walk-forward testing routine conjointly referred to as walk-forward testing, walk-forward validation and a neighborhood of walk-forward analysis.

**ADVANTAGES**

1. Empirically add value to the prediction.

2. Give better results.

## 5. PROPOSED SYSTEM ARCHITECTURE

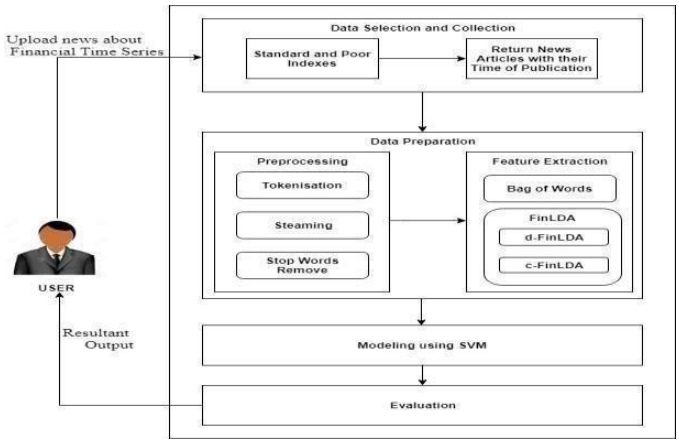


**Figure1:-System architecture**

## 6. ALGORITHM

a. LDA:

Compute the d-dimensional mean vectors for the different classes from the dataset.

1. Compute the scatter matrices.

2. Compute the eigenvectors(e1,e2,...,ed)and corresponding eigenvalues (λ1,λ2,...,λd) for the scatter matrices.

3. Sort the eigenvectors by decreasing eigenvalues.

4.Choose k eigenvectors with the largest eigenvalues to form a d×k dimensional matrix W.

5.Use this d×k eigenvector matrix to transform the samples onto the new subspace.

6. Summarization by the matrix multiplication:Y=X×W

where X=> n×d-dimensional

matrix representing the n samples.

Y=>Transformed n×k- dimensional samples in the new subspace..

b. SVM

It should separate the two classes A and B very well so that the function defined by:

f(x) = a.x + b is positive if and only if

x E A

f(x) 0 if and only if x E B

It exists as far away as possible from all the observations (robustness of the model). Given that the distance from an observation x to the hyperplane is a.x+ b/a.

The width of the space between observations is 2/ a. It is called margin and it should be largest.

Hyperplane depends on support points called the closest points.

Generalization capacity of SVM increases as the number of support points decreases.

## 7. FUTURE SCOPE AND CONCLUSION

The Proposed System introduced FinLDA to extract higher options from news articles for the prediction. This FinLDA is AN extension of the Latent Dirichlet Allocation model that takes changes in monetary statistic into consideration. The extracted options are often employed in any machine learning rule to predict monetary results. In our experiment, parameters of our 2 FinLDA variants (one with separate knowledge and therefore the different with continuous variables describing changes) were calculable by mistreatment each news articles from Reuters and normal & Poor's five hundred Index data and therefore the final outputs from {the 2|the 2} FinLDA variants were used as input options in two standard machine learning algorithms, i.e., Sluzhba Vneshney Razvedki and BPNN, to validate the advantage of the options from FinLDA once examination with different options. The main objective of this research is to predict the stock movement based on the contents of relevant news articles which can be accomplished by building a prediction model.

## ACKNOWLEDGEMENT

Predicting the behaviors of the stock market is always an interesting topic for not only financial investors, but also scholars and professionals from different fields, because successful prediction can help investors to give up major profits. Previous researchers have shown the strong connection between financial news and their impacts to the movements of stock prices. This paper proposes an approach of using time series analysis and text mining techniques to predict daily stock market trends.

## REFERENCES

1) E. F. Fama, ``Efficient capital markets: A review of theory and empirical work,'' *J. Finance*, vol. 25, no. 2, pp. 383_417, May 1970.

2) E. F. Fama, ``Efficient capital markets: II,''*J. Finance*, vol. 46, no. 5, pp. 1575_1617, 1991.

3) R. J. Shiller, ``Speculative prices and popular models,'' *J. Econ. Perspect.*, vol.4,no. 2, pp. 55_65, 1990.

4) L. Blume, D. Easley, and M. O'Hara, ``Market statistics and technical analysis: The role of volume,'' *J. Finance*, vol. 49, no. 1, pp. 153_181, 1994.

5) M. T. Leung, H. Daouk, and A.-S. Chen, ``Forecasting stock indices: A comparison of classification and level estimation models,'' *Int. J. Forecasting*, vol. 16, no. 2, pp. 173_190, 2000.

6) H. Ince and T. B. Trafalis, ``Kernel principal component analysis and support vector machines for stock price prediction,'' in *Proc. IEEE Int. Joint Conf. Neural Netw.*, vol. 3, Jul. 2004, pp. 2053_2058. doi: 10.1109/IJCNN.2004.1380933.

7) R. S. Tsay, *Analysis of Financial Time Series* (Series in Probability and Statistics), 2nd ed. Hoboken, NJ, USA: Wiley, 2005.

8) C.-F. Tsai and Y.-C. Hsiao, ``Combining multiple feature selection methods for stock prediction: Union, intersection, and multi- intersection approaches,'' *Decis. Support Syst.* vol. 50, no. 1, pp. 258_269, 2010.

9) B. Wüthrich, D. Permunetilleke, S. Leung,V. Cho, J. Zhang, and W. Lam, ``Daily prediction of major stock indices from textual www data,'' in *Proc. 4th Int. Conf. Knowl. Discovery Data Mining (KDD)*, 1998, pp. 364_368.Available: http://dl.acm.org/citation.cfm?id=3000292.3 000361

10) W. S. Chan, ``Stock price reaction to news and no-news: Drift and reversal after headlines,'' *J. Financial Econ.*, vol. 70, no. 2, pp. 223_260,2003

11) W. F. M. De Bondt and R. Thaler, ``Does the stock market overreact?'' *J. Finance*, vol. 40, no. 3, pp. 793_805, Jul. 1985. [Online]. Available: http://www.jstor.org /stable/2327804

12) R. J. Shiller, ``From efficient markets theory to behavioral finance,'' *J. Econ. Perspect.*, vol. 17 no. 1, pp. 83_104,2003

13) G. P. C. Fung, J. X. Yu, and W. Lam, ``News sensitive stock trend prediction,'' in *Proc. Pacific-Asia Conf.*

*Knowl. Discovery Data Mining*. Berlin, Germany: Springer-Verlag,2002

14) K. S. Jones, ``A statistical interpretation of term specificity and its application in retrieval,'' *J. Document.*, vol. 28, no. 1, pp.11