2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT)

A Comprehensive Survey on Sales Forecasting Models Using Machine Learning Algorithms

Rameshwaram Sai Mallik
Computer Science Engineering(Artifical Intelligence and Machine Learning) Department
Vardhaman Collage of Engineering
Hyderabad, India
rameshwaramsaimallik@gmail.com

R Abhiram
Computer Science Engineering(Artifical Intelligence and Machine Learning) Department
Vardhaman Collage of Engineering
Hyderabad, India
rathnamalaabhiram9803@gmail.com

Seguru Rithvik Reddy
Computer Science Engineering(Artifical Intelligence and Machine Learning) Department
Vardhaman Collage of Engineering
Hyderabad, India
segururithivikreddy20aiml@vardhaman.org

Jagadish R M
Computer Science Engineering (Artifical Intelligence and Machine Learning) Department
Vardhaman Collage of Engineering,
Hyderabad,India
rm.jagadish@gmail.com

2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT) | 978-1-6654-5635-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICERECT56837.2022.10060168

***Abstract*— Sales Forecasting is a most commonly used in marketing. Nowadays a large number of companies are using this technique to manufacture their product. In this we are going to study about the usage of different Machine Learning Models and Techniques used for sales prediction. The overall study of models and techniques is to increase the efficiency of future sales prediction. Nowadays for any product there are lakhs of reviews were generated by the users on different products in the market. Which confuses customers to make decision whether to buy product or not. And for a specific company to study overall reviews is hard to make product manufacture. This study mainly deals with arranging the opinions of different customers and different kinds of techniques used in sales forecasting. The present work uses mainly four machine learning algorithms namely Support Vector Machine (SVM), Decision Tree (DT), Linear Regression, Random Forest, and K-Nearest Neighbors, K-means Clustering, Logistic Regression for classifying reviews. The forecasting accuracy of each algorithm is evaluated with the Root Mean Square Error (RMSE). The study found that Random Forest is the best model because it had lowest Root Mean Square Error (RMSE) compared to other model and for classifying reviews the Logistic Regression is giving accurate result.**

Keywords-- Prediction, Forecasting System, Machine Learning, Sales Forecasting ,Opinion Mining

I. *Introduction*

A. The main goal of every company or any retail store is to get maximum benefit. This is possible only when they sell more quantity of products, and as a result they gain maximum income. The manager of the particular company or retail store plays an important role to increase the deals of their products by estimating the future sales and arranging the materials, workers and staffs. One of the most important information a company or any store can have the information created by the users, customers, clients. Based on the information given by the users a companies or stores is going find the patterns using machine learning algorithm and this can give raise to a more efficiency forecasting of sales. Managers should try to build good network and make new habits that will improve their performance and to meet customer satisfaction. Few of the techniques are provided by machine learning to give solutions to any kind of complex issues which are hard to address. The Sales forecasting is the basic for production and to get profit. In the areas like where the products are used quickly, sales estimation becomes difficult.

B. Some of the customers who buys items like rice, vegetables, milk etc. have short time of usability and this kind of products can be selected by the customers easily. When it comes to some products like electronic products, clothes which have very short lifecycle, products which can be outdated so quickly etc. Customers will face difficulty while selecting the product. Mainly sales forecasting is used when a retail industry or company want to make a profit without wastage raw material and to make the product cost affordable to everyone. Nowadays many stores, or any kind of industries don't have any kind of clarity on future sales. This is mainly because of lack of knowledge on the products, information to estimate sales. There are some techniques to predict sales in grocery stores depends on some models. The use of the models for forecasting sales results in kind of difficulties like wastage of products, loss in business etc. This type of models generally produced a result in a poor performance. However, machine learning has become the important subject that has produced a significant result in forecasting with maximum efficiency. To accurately estimate future sales, a machine learning is going to follow supervised learning. In this learning a model is built based which can learn based on the previous data history and going to determine the prediction which is required for the client. To accurately estimate future sales, a machine learning model is going trained using the past information by the clients, users from which it is going to predict the future sales. An efficient forecasting model can be used for any company or industry income and to make profits and also provides some additional information about the nature and type of the clients who make better profit in business.

C. Nowadays the online shopping has become famous. To buy any kind of items including vegetables and electronic items the people are showing interest to buy them in online

978-1-6654-5635-7/22/$31.00 ©2022 IEEE

shopping sites. Dealers are frequently requesting their product clients and users to share their reviews on the products they purchased. By this millions of reviews created by the clients and users across the Internet. This had made the internet a major source for getting ideas on the products. As the reviews given by the customer can be negative or positive depends on their experience. However, there are millions of reviews positive and negative feedbacks on a single product which makes the new purchase confuse whether he should take that product or not. Analyzing the feedbacks is mandatory for all companies and retail industries. For all online firms inspecting this feedback plays an important role for their business profits. In recent years the machine learning models are playing an important role in review suggestions for customers based on their interest.

The major steps taken for sales forecasting are basically 1) collecting all the required information related to the products 2) pre-processing the information 3) separating the reviews as positive or negative. In addition to this all we need to analyze the customers opinions on products. The conclusion we are getting from this study is the advantages of applying Machine Learning to sales forecasting. And to find out if they are any advantages over the previous or traditional techniques .

D. In section 2, we are going to study about the various literature reviews followed by in section 3, we will observe from some problem definitions. And In section 4, we will discuss about various algorithms and techniques used in sales forecasting and some important data-preprocessing steps required to build machine learning model. In section 5, we are going to discuss about the performances of the model using some prediction algorithms. In the end the result is summarized in conclusion section by analyzing the future scope and efficiency, accuracy of the model.

II. RELATED WORKS

A. Literature Survey -1

The writing exploration performed in this research paper on sales forecasting using machine learning b exploring different important papers related to machine learning models and forecasting techniques. As we came to know the machine learning is used to apply for business sectors, grocery stores, and retail industries etc. And by using old techniques we faced a lot of difficulties on sales prediction and also the old models are using some sort of forecasting techniques which results to poor performances and less efficiency. Soto overcome this drawback a lot of AI models are built which will produce more effective and high precision results on sales forecasting. All AI models are going to learn from the past information and produce the result with maximum accuracy. Sales prediction is playing an important role in today's business world. If the information is not sufficient, missing information, any null values in the information we are going to get a poor result. To overcome this, we are going to replace the anomalies with mean or any constant values.

B. Literature Survey -2

The sales forecasting for quick consuming goods, food storage shops, retail stores, industries etc. are benefited a lot. The Machine Learning models which are built by different machine learning algorithms and techniques are helping the customers and also the industries to make maximum profit. One major advantage noticed is when compared to other old techniques on sales forecasting the machine learning techniques are played a major role and they are more flexible. The machine learning algorithm uses regression techniques to solve maximum sales problems. The research paper [1] explored a stacking method for creating the regression ensemble of single models. The studies suggest that revelatory model performance for sales forecasting can be increased by using stacking method. The research paper [2] explored that relative mean error (MAE) is the most commonly used techniques for error estimation which is given by error=MAE/mean(sales) 100%. The commonly used techniques in sales forecasting are weighted moving average strategies. The most commonly used algorithm for sales forecasting is random forest algorithm which gives maximum accuracy. The accuracy of the model is highly dependent on the customers, area of the implementation. The commonly used Root-Mean Square Error (RMSE) is used to predict the accuracy of the model. The lower the RMSE value, leads to higher the efficiency of the model. The examination [3] utilizes as an exploration approach to build up a house price expectation model. To improve the performances of this model they used 5369 houses in Fairfax country and build up a housing price forecasting model which depends on Artificial Intelligence (AI) calculations like C4.5, RIPPER, Navies Bayesian. The research showed that the RIPPER algorithm showed maximum accuracy and beats all the old models. In this work [4], most relevant demand framework is created. The updated model mostly depends on historical information by using different machine learning algorithms. The given framework is a combination of nine different time series techniques including exponential smoothing, regression models, moving average (MA) including some multilayer feedforward artificial neural network (MLFANN). The results showed that the given frame work showed maximum results with improved efficiency than old frameworks.

In the paper [5], sales forecasting deals three machine learning algorithms namely Gradient Boosting, Random Forest, K-Nearest-Neighbor. Out of these three models the Gradient Boosting is performing very efficiently by showing maximum accuracy. The main drawback of Gradient-Boosted model is undergoing the overfitting to the dataset. And K-Nearest-Neighbor is showing least performance. And by observing the model's performances we came to know that the more the information for a specific model the more the accuracy it is going to show while predicting the result.

II. PROBLEM IDENTIFICATION

The manager of any company or retail stores like supermarket etc. plays an important role to forecast the sales and identify the demand of the products in the future and to manage the old stock followed by ordering the new stock according to future demands and he should take care of the man power in the stores.

The major problems found are:

- Low accuracy in existing forecasting models.
- Lack of required information or present of anomalies in the data.

- Inefficiency of statical methods to manage a large data.
- poor performances of some models.

The main aim of the project is to satisfy three things:

- To highlight the machine learning algorithms that are used in sales forecasting over the traditional methods.
- The methods and techniques that are followed to classify the comments on a particular product.
- To make products affordable to the customer and also to make profit to store or the company.

III. METHODS

Sales Forecasting is a method of predicting the future sales of a particular product. This is achieved by observing the trend and categorizing the reviews of the customers. There are so many traditional methods that are used in order to predict the future sales of product. But Machine Learning sales forecasting has made great improvement in the sales forecasting. This Machine Learning forecasting techniques uses a large number of machine learning algorithms and also it should have large amount of data to predict the future sales. The detailed System Architecture is shown in the fig.1, Initially the raw data is converted in a data set by organizing the data followed by dividing the data into two parts 1) Training data 2) Testing data. The training data set should be 80% of the original data and remaining data, 20% is taken as testing data. Using some machine learning algorithms and training data a sales forecasting model is build. Using testing data we will going to predict the accuracy of the sales prediction.

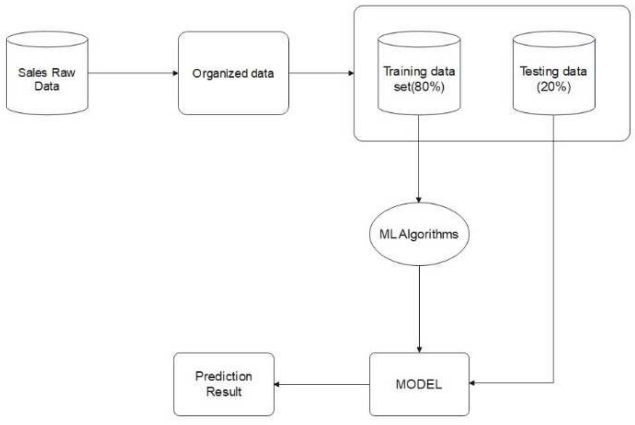

Fig.1 System Architecture

A. Sales Prediction Algorithms

1) Random Forest

Random Forest algorithm is mostly commonly used algorithm in machine learning for prediction the future goals. It is easy to use and gives maximum accuracy in most of the results. It is easy to use and gives maximum accuracy in most of the results. Basically, Random Forest is the combination of large number of Decision Trees algorithms which produce more accuracy. The other name of Random Forest is Ensemble Learning. The Random Forest depends on the number of trees involved in it. The more the number of the trees included the more will be the accuracy of the algorithm. The main drawback of Decision Tree is overfitting which results in maximum of accuracy in training data but low accuracy in testing data. Random Forest takes advantage of this drawback and allows decision tree to access the data to produce the different Decision Trees.

2) Decision Tree

The Decision Tree algorithm mostly used to solve problems related to regression and classification. The problems are solved by representing the given data set in the form of nodes and vertices. Where the nodes represent attributes and vertices representing the rules to be applied followed by child nodes shows the result. This algorithm compares the values of root node with all other attributes and moves to its next nodes. Before going to other the comparison of attributes values is going to take place. This process continues until it reaches to leaf nodes.

3) Support Vector Machine (SVM)

The Support Vector Machine is one most popular algorithm which is used for both classification as well as regression. Most commonly it is used for classification problems. It consists of three components namely positive hyperplane, negative hyperplane, support vectors. The distance between positive and negative hyperplane is called Maximum margin. The larger the distance between positive and negative hyperplane the higher the accuracy of the model. This SVM is divided into two types namely Linear SVM and Non-Linear SVM. In simple linear SVM is used for linearly separable data, where as Non-Linear SVM is used for non-linear separated data.

4) Linear Regression

Linear Regression is an algorithm that deals with problems related Regression. When attribute values are not change able then that form is called as regression problem. A linear relationship is formed between the dependent and independent variable in order to get the result. In this model creates a mathematical function that produce a straight line's which meets all the data points. Each line consists of different accuracy and error the line which shows maximum efficiency in prediction is decided as best fit line. The commonly used equation line for linear regression is Y=a + b X, where X independent variable and Y is dependent variable. The Line slope is 'b' and intercept is 'a'.

B. Sales Forecasting Techniques

1) Delphi method

This method is also called as estimate-talk-estimate techniques (ETE). This is a technique which involves the gathering of information from the experts through asking several questions. This method mainly depends on the experts who is having a knowledge on their domain. So they can predict the outcome of the future scenario.

2) Time series analysis and projection

This is the process of analyzing the information collected over the specific intervals of time. In this the analyst tries to record the information in a specific interval of time not considering the random information. This requires a large number of data points to predict accurately. This analysis mainly used to find the systematic pattern in the trends or life style over time.

3) weighted moving average

The weighted moving is one of the most popular and best technique. Nowadays most companies are using weighted moving average to find the direction of trend. As it will assign the higher weights to recent data points and lesser weights to old data points, by this it will differentiate the information to make prediction more accuracy. The formula used to calculate the moving average M= sigma{w(t)*v/w(t)} where t lies in (1-n).

w= weighting factor; v= actual value ; n= number of periods in weighting group.

C. Data Collection

The most important step in designing a machine learning model is Collecting the data. Initially the data collected will be in the raw form. It should be abstracted into an organized form which is easily understandable by the developer. In data set each column represent a variable and reach roe represent a particular data set as shown in Figure 2. Any Machine Learning Algorithms are highly dependent on the data. The more the information/data without anomalies and null values the more will be the accuracy of the model. Planning of data is takes more time which involves the converting data into a more organized form to make it suitable for machine learning applications.

D. Data Pre-Processing

This is the step where the required data is collected after organizing it into a particular format. The Pre-Processing mainly involves the gathering of required from the overall data collection and making a data set called as training data set which is used to train the particular machine learning model followed by creating a testing data set which is used to predict the accuracy of the model as shown in Figure 2. Mostly used ratio to divided the data collection into training and testing data is 80:20. This data pre-processing is the most critical step in building any machine learning model.

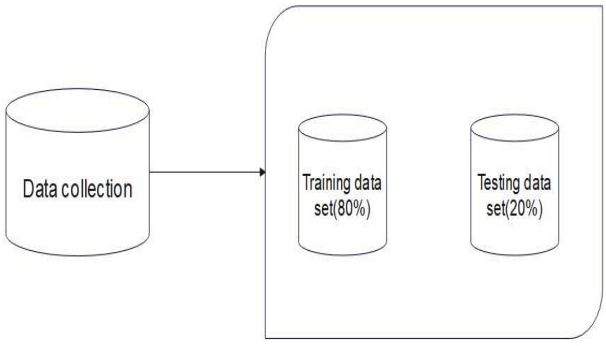

Fig.2 Data Pre-Processing

E. Dealing with Missing values and Null values

It is most common nowadays that any data set can contains missing values and null values. The Machine Learning algorithms may show low accuracy due to the presence of this missing and null values. Dealing with missing values plays an important role to gain maximum accuracy for any machine learning model. There are certain methods to overcome the drawbacks of missing and null values:

1) Replacing the missing value with any constant value.

2) Replacing the missing value with mean, median or mode

.

F. Prediction

This is the step where perform the testing of our model to determine the accuracy and performance of the model. The testing data set is used to test the performance of the model. Initially we use training data i.e. Historical or past data is used to train the model followed by training data set is used to test the performance or prediction result of the model. Prediction plays an important role to know the accuracy of the model. Many industries use's this technique to know whether their model is predicting the required result or not.

G. Opinion Mining

This is mainly used categorize the given sentences or reviews on a particular product into positive or negative. This process is mainly deals with the concept of different neural networks i.e. Artificial neural network (ANN), Convolutional neural network (CNN). Initially the sentence which is to be analyzed followed by feature selection process is going to takes place where the particular feature is selected to find the polarity of the sentence. The machine learning algorithms such as K-Nearest Neighbor (KNN), K-means clustering, Logistic Regression. Out of this Three Logistic Regression is best algorithms for classifying the reviews.

H. Logistic Regression

The most popular technique used for classification process is Logistic Regression. The process involves the estimating the value of the single variable with the help of two or more independent variables. The result of this classification algorithm will be in the form of binary result which gives the output only in two formats

1) True -Positive Review

2) False – Negative Review.

The overall process of logistic Regression is represented in Figure 3. Initially a Threshold value is set to predict in which class a review belongs to.

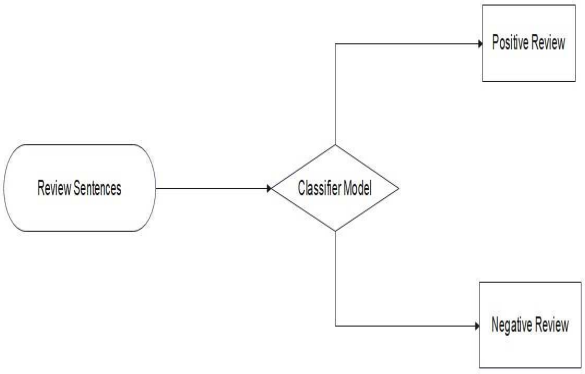

Fig.3 Logistic Classifier

I. *Loading of Data*

It is important to understand whether the product review is good or bad. That can be possible by loading the history of feedbacks and trying to preform data preprocessing steps to divided the data into testing and training. Then using the training data set we can build the model with logistic regression algorithm which can decided the feedback whether it is positive or negative based on the study of previous feedback. So we should build a word dictionary to help the model to understand the feedbacks.

J. *Analysis of Data*

This step plays an important role to classify the review into positive or negative. In this process we are going to use the Logistic Regression Algorithm to get the result. In sentiment analysis the model is going to learn which is positive and negative based on training. The fixed values are decided to determine whether the given comment is negative or positive. In most cases 1 is fixed for positive value and 0 is fixed for negative value. The feedbacks from the training data will be used to build a ML model by analyzing the feedback and tries to predict which is positive and negative feedbacks.

IV. RESULTS AND DISCUSSION

In this section the overall result of the sales forecasting is summarized. The accuracy percentage represents overall correct prediction of the values from the test data set by the Machine Learning Model.

Sales Prediction:

The model prediction is the last and the important step to analyze the performance of the model and to determine the accuracy of the prediction of future sales. The most commonly used technique to find the error in any machine learning model is Root Mean Square Error (RMSE). According to this technique it is going to calculate the square root of the mean of all the errors obtained for n number of inputs and it is taken as the Error Rate. The Figure 4 shows the error rate of different machine learning algorithms. By observing the graph we can came to know that the best machine learning algorithm which is used for sales prediction is Random Forest Algorithm as it is showing lowest root mean square error i.e. (9%) compared to all other algorithms.

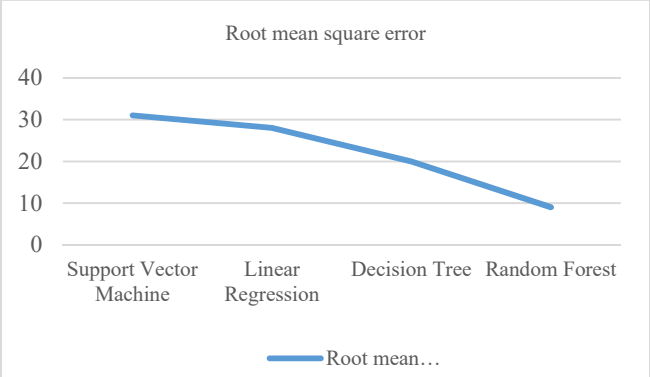

Fig.4 Percentages of Root Mean Square Error in Different Algorithms

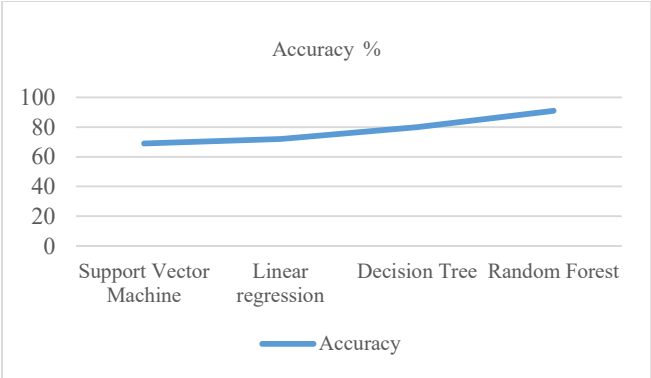

Fig.5 Accuracy representation of all Algorithms.

As shown in figure 5 the Random Forest Algorithm is showing more accuracy compared to other possible algorithms. As shown in figure 4 the error percentage using random forest is 9% which is low compared to other algorithms so Random Forest Algorithm is showing more accuracy in sales prediction which is nearly 91%.

V. CONCLUSION

Finally, the researchers conclude that to handle a large amount of data like comments, reviews on products the business companies or retail Industries requires machine learning models to predict the future sales of a product. Using sales forecasting models different retail industries, supermarket got benefited. For building sales forecasting model four main machine learning algorithms were used namely 1) Support Vector Machine (SVM) 2) Linear regression 3) Decision Tree 4) Random Forest. Out of this four Random Forest showed the better performance with an accuracy of 91% compared to all other algorithms besides the it had lowest root mean square error compared to all other algorithms. And also the performances of a model will increases if we have large amount of data.

Till now there are large number of practices are going on for review classification. This paper explored a general process for review classification. Opining mining is one of the Technique that is used to classify the comments, reviews into positive and negative. Nowadays the online shopping is becoming very popular. As technology is growing a large number of users or customer are trying to purchase their products online. Every product contains a huge number of reviews or comments and on the basis of these reviews the buyer is going to purchase the product. As there are lakhs of reviews on a particular product the buyer unable to decide whether to buy a particular product or not. To overcome this problem Logistic Regression algorithm is used to classify the reviews into positive and negative and helps the customer to purchase the right product at affordable price which is benefit for customers as well as company.

In this paper explored four different algorithms to build sales forecasting model. Based on the result concluded that random forest is best algorithm for sales forecasting. Nowadays large number of companies are using this sales forecasting model to predict the future sales and they also deciding the price which is affordable by all types of customers besides the price is also suitable to make a good

profit for the companies. In today's business world this sales forecasting techniques are used in essential area.

REFERENCES

[1] Machine Learning, Tom Mitchell, McGraw Hill,1997.

[2] Large-Scale Video Classification with Convolutional Neural Networks , by Fei-Fei, L., Karpathy, A., Leung, T., Shetty, S., Sukthankar, R., & Toderici, G. (2014).

[3] Learning deep features for scene recognition using places database , by Lapedriza, À., Oliva, A., Torralba, A., Xiao, J., & Zhou, B. (2014). NIPS.

[4] Park, B., & Bae, J. k. (2015) Using machine L=learning algorithms for housing price prediction: The case of Fairfax country, Virginia housing data. ELSEVIER – Expert Systems with Applications.

[5] A survey on feature selection methods, by Chandrashekar, G., & Sahin, F. Int. J. on Computers & Electrical Engineering.

[6] Deep Residual Learning for Image Recognition, by He, K., Ren, S., Sun, J., & Zhang, X. (2016). CoRR, abs/1512.03385.

[7] Ask less – Scale Market Research without Annoying Your Customers by Venkatesh Umaashankar and Girish Shanmugam S

[8] Pavlshenko, B.M. (2019). Machine-Learning models for sale time series forecasting.

[9] Sepideh Paknejad (2018). Sentiment classification on Amazon reviews using machine learning approaches.

[10] Kilimci, Z. H., Akyuz, A. O., Uysal, M., Akyokus, S., Uysal, M, O., Atak Bulbul, B., & Ekmis, M.A (2019). An improved demand forecasting model using deep learning approach and proposed decision integration strategy for supply chain. Complexity, 2019.

[11] Rising Odegua (2020) Applied Machine Learning for Supermarket Sales prediction. Research gate

[12] Aneesh Tony, Pradeep Kumar, Rohith Jefferson and Subramanian (2021). A Study and Sales Forecasting Model using Machine Learning Algorithm.

[13] Unsupervised Learning of Probably Symmetric Deformable 3D Objects from Images in the Wild, by Shangzhe Wu, Christian Rupprecht, Andrea Vedaldi

[14] Targeted Aspect-Based Sentiment Analysis via Embedding Commonsense Knowledge into an Attentive LSTM by Yukun Ma, Haiyun Peng, Erik Cambria

[15] A Deep Probabilistic Model for Customer Lifetime Value Prediction, by Xiaojing Wang, Tianqi Liu, Jingang Miao

[16] Context-aware Embedding for Targeted Aspect-based Sentiment Analysis, by Bin Liang, Jiachen Du, Ruifeng Xu, Binyang Li, Hejiao Huang