Product Review Classification by using Convolution Neural Networks and KNN Classifier

Gorti Satyanarayana Murty[1], K Amit[2], H.Vineetha[3]
Department of Computer Science and Engineering
Aditya Institute of Technology and Management

Abstract- Online shopping websites usually ask their customers to review the products and associated services. As ecommerce is becoming more and more popular, the number of customer reviews grows rapidly. It is difficult for a potential customer to read all original reviews in order to make a decision on whether to buy the product or not. So product reviews classification has grown to be one of the hottest research areas in sentiment analysis. To solve this issue, e-commerce websites provide reviews classification. We use two methods to solve this problem machine learning and deep learning algorithm: K-Nearest Neighbor (KNN) and convolution neural networks and compares their overall accuracy. This paper elaborately discusses product review classification based on long short-memory (LSTM) recurrent neural network and K-NN classifier.

Keywords- sentiment analysis; deep learning; recurent nueral network; long-short memory; KNN.

I. INTRODUCTION

With the rapid popularization and development of network, the network information resources which are in explosive growth. The users place their sentiments on the web, especially on the social network and online shopping websites. People share their opinions on the websites and show their reviews for a product. When we need to choose a product, we always refer the reviews of others. So these reviews are valuable to individual consumers for making correct decisions and business organizations for improving their products. On the other hand, the rapid growth of unstructured data coincides with the burst of social media and e-commerce, new opportunities and challenges arise as people now can use the huge volume of unstructured data to mining meaningful information. For these reasons, product review classification, one of the sentiment analysis (SA) tasks, has received a sharp boost of research interest in natural language processing (NLP). In fact, sentiment analysis also promotes the development of machine learning (ML) due to many machine learning methods have been used in this area[1]. There are mainly three basic tasks in a typical application of sentimental analysis: namely holder detection, target extraction and sentiment classification. Sentiment classification is the most important and representational application. Product reviews classification is the branch of sentiment classification. It can be divided into three levels: document level , sentence level , entity and aspect level[1,7,8] .

II. LITERATURE SURVEY

A.MACHINE LEARNING

Machine learning is one of the fastest growing areas of computer science, with far-reaching applications. It refers to the automated detection of meaningful patterns in data. Machine learning tools are concerned with endowing programs with the ability to learn and adapt. Machine Learning has become one of the mainstays of Information Technology and with that, a rather central, albeit usually hidden, part of our life. With the ever increasing amounts of data becoming available there is a good reason to believe that smart data analysis will become even more pervasive as a necessary ingredient for technological progress.

There are several applications for Machine Learning (ML), the most significant of which is data mining. People are often prone to making mistakes during analyses or, possibly, when trying to establish relationships between multiple features .

Data Mining and Machine Learning are Siamese twins from which several insights can be derived through proper learning algorithms. There has been tremendous progress in data mining and machine learning as a result of evolution of smart and Nano technology which brought about curiosity in finding hidden patterns in data to derive value. The fusion of statistics, machine learning, information theory, and computing has created a solid science, with a firm mathematical base, and with very powerful tools.

Machine learning algorithms are organized into a taxonomy based on the desired outcome of the algorithm. Supervised learning generates a function that maps inputs to desired outputs. Unprecedented data generation has made machine learning techniques become sophisticated from time to time. This has called for utilization for several algorithms for both supervised and unsupervised machine learning. Supervised learning is fairly common in classification problems because the goal is often to get the computer to learn a classification system that we have created. Machine learning algorithms are classified into three categories:

- Supervised Learning – Train Me!
- Unsupervised Learning – I am self sufficient in learning
- Reinforcement Learning – My life My rules! (Hit & Trial)

1. Supervised learning:

Supervised Learning is the method, wherein the training data includes both the input and the desired results. Training the system with examples is called supervised learning. Or else, training the algorithm with a teacher can also be treated as supervised learning. After training the algorithm with all sample data or labelled data, which has both the predictors on the target variable, one can train the algorithm and use the unseen example for further classification.

Here are some of the important features of Supervised Learning in Mahout:

- The construction of a proper training, validation and test set (Bok) is crucial.
- These methods are usually fast and accurate.
- The Supervised Learning methods have to be able to generalize.

They give correct results, when new data are given in input without knowing a priori target.In some cases, the correct results (targets) are known and given in input to the model during the learning process.

2. Unsupervised learning:

The model learns through observation and finds structures in the data. Once the model is given a dataset, it automatically finds patterns and relationships in the dataset by creating clusters in it. What it cannot do is add labels to the cluster, like it cannot say this a group of apples or mangoes, but it will separate all the apples from mangoes.

Suppose we presented images of apples, bananas and mangoes to the model, so what it does, based on some patterns and relationships it creates clusters and divides the dataset into those clusters. Now if a new data is fed to the model, it adds it to one of the created clusters.

3. Reinforcement learning:

It is the ability of an agent to interact with the environment and find out what is the best outcome. It follows the concept of hit and trial method. The agent is rewarded or penalized with a point for a correct or a wrong answer, and on the basis of the positive reward points gained the model trains itself. And again once trained it gets ready to predict the new data presented to it [10, 11].

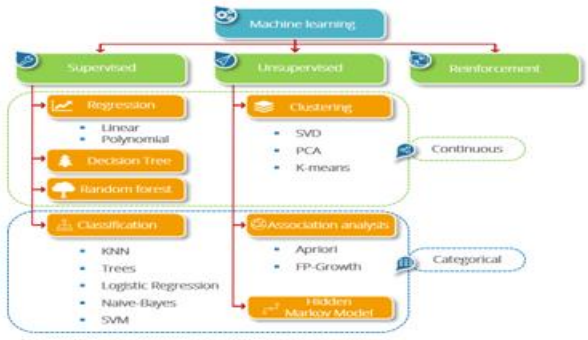

Fig.1: Types of machine learning

4. Supervised machine learning algorithms:

a. K-Means:

K-means: According to and Kmeans is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori.K- Means algorithm is be employed when labeled data is not available .General method of converting rough rules of thumb into highly accurate prediction rule. Given ―weak‖ learning algorithm that can consistently find classifiers (―rules of thumb‖) at least slightly better than random, say, accuracy _ 55%, with sufficient data, a boosting algorithm can provably construct single classifier with very high accuracy, say, 99%.

b.Naive Bayesian:

Naive Bayesian (NB) Networks: These are very simple Bayesian networks which are composed of directed acyclic graphs with only one parent (representing the unobserved node) and several children (corresponding to observed nodes) with a strong assumption of independence among child nodes in the context of their parent .Thus, the independence model (Naive Bayes) is based on estimating . Bayes classifiers are usually less accurate that other more sophisticated learning algorithms (such as ANNs).However, performed a large-scale comparison of the naive Bayes classifier with state-of-the-art algorithms for decision tree induction, instance-based learning, and rule induction on standard benchmark datasets, and found it to be sometimes superior to the other learning schemes, even on datasets with substantial feature dependencies. Bayes classifier has attribute independence problem which was addressed with Averaged One-Dependence Estimators.

c.Support Vector Machines (SVMs):

These are the most recent supervised machine learning technique .Support Vector Machine (SVM) models are closely related to classical multilayer perceptron neural networks .SVMs revolve around the notion of a ―margin‖—either side of a hyperplane that separates two data classes. Maximizing the margin and thereby creating the largest possible distance between the separating hyper plane and the instances on either side of it has been proven to reduce an upper bound on the expected generalization error.

d. Decision Trees:

Decision Trees (DT) are trees that classify instances by sorting them based on feature values. Each node in a decision tree represents a feature in an instance to be classified, and each branch represents a value that the node can assume. Instances are classified starting at the root node and sorted based on their feature values. Decision tree learning, used in data mining and machine learning, uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. More descriptive names for such tree models are classification trees or regression trees .Decision tree classifiers usually employ post-pruning techniques that evaluate the performance of decision trees, as they are pruned by using a validation set. Any node can be removed and

assigned the most common class of the training instances that are sorted to it.

5. LIMITATIONS OF MACHINE LEARNING

Machine Learning is not capable of handling high dimensional data that is where input & output is quite large. Handling and processing such type of data becomes very complex and resource exhaustive. This is termed as Curse of Dimensionality.

1 DEEP LEARNING

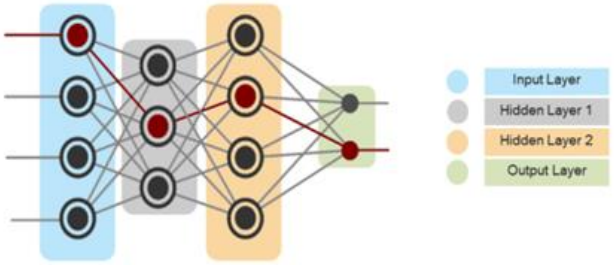

Fig.2: Architecture for deep neural network

Any Deep neural network will consist of three types of layers:

- The Input Layer
- The Hidden Layer
- The Output Layer

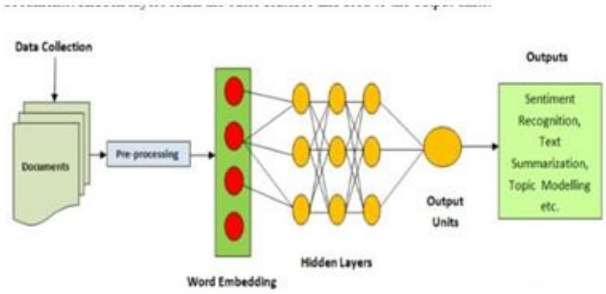

Fig.3: Architecture for deep neural network for text analysis

In the above diagram, the first layer is the input layer which receives all the inputs andthe last layer is the output layer which provides the desired output. All the layers in between these layers are called hidden layers. There can be n number of hidden layers thanks to the high end resources available these days.

The number of hidden layers and the number of perceptions in each layer will entirelydepend on the use-case you are trying to solve.

Now that you have a picture of a Deep Neural Networks, let's move ahead in this Deep Learning Tutorial to get a high level view of how Deep Neural Networks solves a problem of Image Recognition [3,5,6].

2. CONVOLUTIONAL NEURAL NETWORK

Convolution neural network is one of the feed-forward artificial neural network models for deep learning. The connectivity pattern between its neurons is inspired by the organization of the animal visual cortex. CNN has three important characteristics, local connectively, parameter sharing and down-sampling so that CNN is also a neural network model with shift invariant and space invariant. CNN can be considered including two main parts. The first part plays a role in feature extraction and the second part can be thought as a classifier. The first part contains multi convolution and pooling layers. A complex network has multiple convolution-pooling pairs. Each layer takes the output of its previous layer as the input. Convolutional layer is the core of a CNN. Traditional Neural Network (NN) has three layers - one input layer, one hidden layer and one output layer with activation function. Deep neural networks are modified version neural networks with multiple hidden layers. In deep neural network, hidden layers play a major role in learning the features from input data sets. The layers of CNN are

1) Pooling

Convolution networks may include local or global pooling layers which combine the outputs of neuron clusters at one layer into a single neuron in the next layer. For example, *max pooling* uses the maximum value from each of a cluster of neurons at the prior layer. Another example is *average pooling*, which uses the average value from each of a cluster of neurons at the prior layer.

2) Fully connected

Fully connected layers connect every neuron in one layer to every neuron in another layer. It is in principle the same as the traditional multi-layer perceptron neural network (MLP). The flattened matrix goes through a fully connected layer to classify the images.

The parameters of a convolutional layer consist of a set of learnable filters or kernels with a small receptive field named local connectively. Convolutional layer has many convolution kernels which have different weights. Each kernel is connected to only a small region of the input volume, but extends along entire depth of the input volume through sliding window. This architecture ensures that the productions of learnt filters are the strongest response to a spatially
local input pattern. Parameter sharing means weights invariant in window sliding process. Another important concept of CNN is pooling which is a form of nonlinear. Max-pooling is the most common functions to implement pooling. It partitions the output of convolutional layer into a set of non-overlapping rectangles and, for each such sub-region, outputs the maximum. Pooling operation provides a form of translation invariance and reduces the spatial size of the features. This mechanism can decrease the amount of parameters and reduce computational complexity of the network, and hence to control over fitting. The second part is the fully-connected layer and loss layer which is a typical feed forward neural network. It can add one or more hidden layers to fit features by supervised learning. CNN is trained by using back-propagation algorithm and all parameters are jointly optimized [5].

K-NN is a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification. It is non parametric

method used for classification or regression. In case of classification the output is class membership (the most prevalent cluster may be returned), the object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its *k nearest neighbors*. This rule simply retains the entire training set during learning and assigns to each query a class represented by the majority label of its k-nearest neighbors in the training set.

The *Nearest Neighbor* rule (NN) is the simplest form of K-NN when K = 1. Given an unknown sample and a training set, all the distances between the unknown sample and all the samples in the training set can be computed. The distance with the smallest value corresponds to the sample in the training set closest to the unknown sample. Therefore, the unknown sample may be classified based on the classification of this nearest neighbour. The K-NN is an easy algorithm to understand and implement, and a powerful tool we have at our disposal for sentiment analysis. K-NN is powerful because it does not assume anything about the data, other than a distance measure can be calculated consistently between two instances . As such, it is called non- parametric or non-linear as it does not assume a functional form. The flowchart of K-NN Classifier Nearest neighbor methods is considered one of the simplest and most yet effective classes of classification algorithms in use. Their principle is based on the assumption that, for a given set of instances in a training set, the class of a new yet unseen occurrence is likely to be that of the majority of its closest "neighbour" instances from the training set. Thus the k-Nearest Neighbour algorithm works by inspecting the k closest instances in the data set to a new occurrence that needs to be classified, and making a prediction based on what classes the majority of the k neighbours belong to. The notion of closeness is formally given by a distance function between two points in the attribute space, specified a priori as a parameter to the algorithm. An example of distance function typically used is the standard Euclidean distance between two points in an n-dimensional space, where n is the number of attributes in the data set [2,4,9].

II. RELATED WORK

The existing system compares the performance accuracy of the KNN classifier and naive baye's product review classification. In this paper we proposed an approach to provide a method for the product review classification by using machine learning algorithm(K-NN) and deep learning algorithm (CNN). By using this approach we find the score of the each review.

1.Algorithm

```
Input: Review data set
    Begin
    Initialize the n-sized labelled hash table and assign the n customer reviews
   Initialize
    pos=0,neg=0,supportpos=0,supportneg=0,score=0
 For each i in N reviews
    do word Analysis
    acquire score of hypothesis
 End for
  For each i in N reviews(X is the word value in database)
 do X analysis
  If(X>score)
Supportpos++
 Review[i]=label of temporary hash
table[i]=0 Else
 Supportneg++
 Review[i]=label of temporary hash table[i]=1
 End for
```

Fig.4: Output Of CNN

III. EXPERIMENTAL EVALUATION

a. RESULTS

Fig.5: Dataset given as Input

Fig.6: Individual Review Score

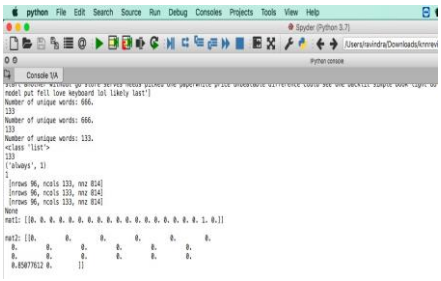

Fig.7: Individual Review Score

b. GRAPHICAL ANALYSIS

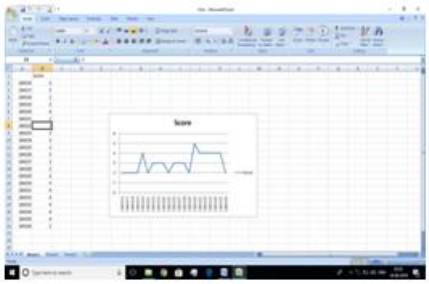

Figure:8 Graphical representation of product reviews score using CNN

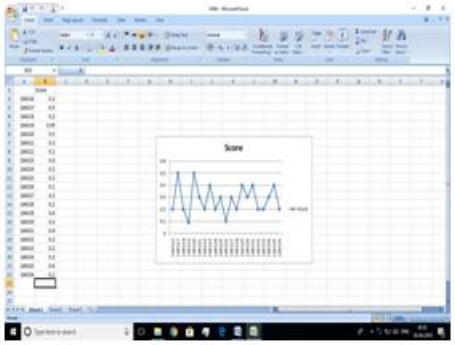

Figure: 9 Graphical representation of product reviews score using K-NN

IV. CONCLUSION

In this we have proposed a method for product review classification by using convolutional neural network and K-Nearest Neighbor algorithm. We compared deep convolutional neural network algorithm and K-NN for sentiment classification of the product reviews of online websites. The experimental results show that the classifiers yielded better results for the product reviews with deep convolutional neural networks the approach giving above more accuracies and outperforming than the k-NN approach. Thus we can say deep convolutional neural networks can be used successfully to analyze product or movie reviews.

V. REFERENCES

[1]. Beyond the Stars: Improving Rating Predictions using Review Text Content, by Gayatree Ganu, No´emie Elhadad, Am´elie Marian on june 28 2009.

[2]. Categorization of 'Holy Quran-Tafseer' using K-Nearest Neighbor Algorithm, by Faris Mahdi Alwan , Geehan S. Hassan on 09 March 2016.

[3]. Keyword extraction from online product reviews based on bi-directional LSTM recurrent neural network, by J. Zhang, Yue Wang on 15 March 2018.

[4]. Sentiment Analysis on Twitter Data using KNN and SVM, by Mohammad Rezwanul Huq, Ahmad Ali, Anika Rahman on November 6 2017.

[5]. Convolutional Neural Networks for Sentiment Classification on Business Reviews, by Andreea Salinca on 14 October 2017.

[6]. Amazon Food Review Classification using Deep Learning and Recommender System, by Lan Xu, Zhenxiang Zhou, December 11, 2016.

[7]. Sentime Datasets, by Jawad Khan, Aftab Alam, , October 2016.

[8]. Sentiment Analysis of Product Reviews, Stephen C.F. Chan, January 2008

[9]. Sentiment Analysis of Review Datasets Using Naïve Bayes and K-NN Classifier, by Lopamudra Dey, Beepa Bose, Sanjay Chakraborty, July 2016.

[10]. Sentiment Analysis Using Deep Learning Technique CNN with KMeans, by B.Swathi Lakshmi, P. Sini Raj and R. Raj Vikram, November 11,2017.

[11]. Sentiment Classification Of Movie Reviews By Supervised Machine Learning Approaches, by P.Kalaivani on Aug-Sep 2013.

Satyanarayana Murty Gorti received his Master's Degree from JNTU Kakinada, AP India , and Ph.D in May 2014 in Computer Science and Engineering and area of Specialization is Image Mining. He is working as a Professor in Aditya Institute of Technology and Management, Tekkali-532201 , AP,India. He is working in this Institute since August 2005 and having 20+ Years of experience in teaching. He acted as a reviewer for various International conferences and Journals. He acted as a Convener for AICTE Sponsored FDPs and Conferences. He published good number of papers in International Journals with good impact factor and Scopus Indexed. He is a Life Member of CSI and ISTE.

I, KENGUVA AMIT, have completed my btech CSE this year (2019 passout) student from ADITYA INSTITUTE OF TECHNOLOGY AND MANAGEMENT.And my achievements are 2nd in poster presentation on recommendation system ASPIRE 2K18 at AITAM in 2018 and successful in Android App development at AITAM in 2017.

I, Vineetha Haridasu , have completed my btech CSE this year (2019 passed out) student from ADITYA INSTITUTE OF TECHNOLOGY AND MANAGEMENT. And my achievements are Getting scholarship from college for acquiring good rank in EAMCET, Done mini project on blood bank management system, Completed one month internship in Flip kart , Hyderabad, and Got selected for two companies cognizant and cyient

.