**Latest News Classifier**

Vedanshu Trivedi1, Umang Unadakat2, Talati Keyur3

1U & PU Patel Dept. of Computer Engineering, Charotar University, Anand, Gujarat

2U & PU Patel Dept. of Computer Engineering, Charotar University, Anand, Gujarat

3U & PU Patel Dept. of Computer Engineering, Charotar University, Anand,Gujarat

**Abstract**

The main goal of classification problem is to identify the category or class to which a new data will be fall under. There are many techniques classify new news comes under which class. From all of those techniques news can be classify about 70% accurately. So, how it can become more accurate that is the issue. From past few years this topic is quite significant. Now a days many sources available to user for data in digital and electronic media. This paper presents types of techniques for classify the latest news. Classify news into categories based on their headline.

Various classifiers were tried - Decision Tree, Support Vector Classifier,Multinomial Naive Bayesian Classifier, Multilayered Perceptron, Random Forest. Multinomial Naive Bayesian Classifier worked the best. It is logical for Multinomial Naive Bayesian to work the best as even we as humans classify based on keywords. We are likely to predict “Politics” is we see keywords lik Obama, election, republic and we are likely to predict “Criminal” if we see keywords like drugs, jail and so on. Naive bayesian scans whole dataset and finds the probabilities of each word in headline being associated with a class and then find the probability for whole headline hence it works good.

1. **Introduction**

Now a days news is easily available due to many resources available on the internet so the number of news are get increase hence it is difficult for user to got the news of their interest so the classification of latest news is require.

For classify news there are many data sources available using those datasets and from available classifying techniques. Class of news can be classified. With the using of those techniques accuracy is around 70%. So this paper is on how it can become more accurate with some modification of dataset or improve techniques. All algorithms are improved by many researches so,

this paper is not on how to improve algorithms. But, improve way of applying algorithms. In step one use this algorithms as it is.

We will select the accurate algorithm among the available algorithm but, for improving accuracy of classifying latest new news we use two-way verification of news on by using available algorithms and second using user’s reviews or feedback.

Types of News: Technology, Sports, Business, Finance, Global, Health, Food, Fashion, Entertainment, Bollywood/Hollywood.

**2.News Classification works**

Right now classifier works using algorithms or techniques.

For improving accuracy one more step added to that algorithm.

**Step 1:**

In first we apply available algorithms to classify the latest news those algorithms are already improves by many researchers so, that in this paper our focus is not on just improving algorithm but improve way of applying algorithms. So we will select the accurate algorithm among the available algorithm

For the Latest News classification first the news is pre-processed for giving them as input for our algorithm for that following steps are followed.



After this pre-processing steps we have to apply any one of the available algorithm. Most commonly known news classification algorithms are Decision Trees, Support Vector Machine, K- Nearest Neighbours, Artificial Neural Network and Naïve Bayes.

**Decision Tree**

It is a classifier for text categorization which represents in a tree foam and leaf node are classes and intermediate nodes represents decision. It uses Recursive Approach.

A decision tree is a flowchart-like tree structure where an internal node represents feature(or attribute), the branch represents a decision rule, and each leaf node represents the outcome. The topmost node in a decision tree is known as the root node. It learns to partition on the basis of the attribute value. It partitions the tree in recursively manner call recursive partitioning. This flowchart-like structure helps you in decision making. It's visualization like a flowchart diagram which easily mimics the human level thinking. That is why decision trees are easy to understand and interpret.



Decision Tree is a white box type of ML algorithm. It shares internal decision-making logic, which is not available in the black box type of algorithms such as Neural Network. Its training time is faster compared to the neural network algorithm. The time complexity of decision trees is a function of the number of records and number of attributes in the given data. The decision tree is a distribution-free or non-parametric method, which does not depend upon probability distribution assumptions. Decision trees can handle high dimensional data with good accuracy.

How does the Decision Tree algorithm work?

The basic idea behind any decision tree algorithm is as follows:

1. Select the best attribute using Attribute Selection Measures(ASM) to split the records.
2. Make that attribute a decision node and breaks the
3. dataset into smaller subsets.
4. Starts tree building by repeating this process recursively for each child until one of the condition will match:

* All the tuples belong to the same attribute value.
* There are no more remaining attributes.
* There are no more instances.

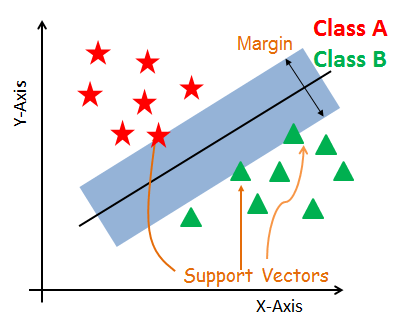
**Support Vector Machine**

It is algorithm mainly used in classification and regression.

In this data is plotted as a point and then from that point class are divide by hyper –plane. It has unique feature that it includes negative and positive training sets.

This algorithm looks for the k points which are near to that vector by distance measure and that k points are nearest neighbours.

Generally, Support Vector Machines is considered to be a classification approach, it but can be employed in both types of classification and regression problems. It can easily handle multiple continuous and categorical variables. SVM constructs a hyperplane in multidimensional space to separate different classes. SVM generates optimal hyperplane in an iterative manner, which is used to minimize an error. The core idea of SVM is to find a maximum marginal hyperplane(MMH) that best divides the dataset into classes.



#### Support Vectors

Support vectors are the data points, which are closest to the hyperplane. These points will define the separating line better by calculating margins. These points are more relevant to the construction of the classifier.

#### Hyperplane

A hyperplane is a decision plane which separates between a set of objects having different class memberships.

#### Margin

A margin is a gap between the two lines on the closest class points. This is calculated as the perpendicular distance from the line to support vectors or closest points. If the margin is larger in between the classes, then it is considered a good margin, a smaller margin is a bad margin.

## How does SVM work?

The main objective is to segregate the given dataset in the best possible way. The distance between the either nearest points is known as the margin. The objective is to select a hyperplane with the maximum possible margin between support vectors in the given dataset. SVM searches for the maximum marginal hyperplane in the following steps:

1. Generate hyperplanes which segregates the classes in the best way. Left-hand side figure showing three hyperplanes black, blue and orange. Here, the blue and orange have higher classification error, but the black is separating the two classes correctly.
2. Select the right hyperplane with the maximum segregation from the either nearest data points as shown in the right-hand side figure.



K Nearest Neighbour

It is one of the simplest algorithm. Here data is represented as vector.

K-Nearest Neighbors, or KNN for short, is one of the simplest machine learning algorithms and is used in a wide array of institutions. KNN is a **non-parametric, lazy**learning algorithm. When we say a technique is non-parametric, it means that it does not make any assumptions about the underlying data. In other words, it makes its selection based off of the proximity to other data points regardless of what feature the numerical values represent. Being a lazylearning algorithmimplies that there is little to no training phase. Therefore, we can immediately classify new data points as they present themselves.

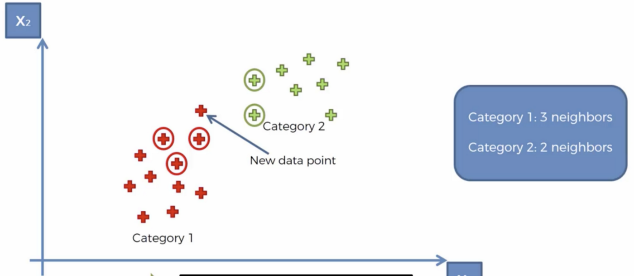
# Some pros and cons of KNN

**Pros**:

* No assumptions about data
* Simple algorithm — easy to understand
* Can be used for classification and regression

**Cons**:

* High memory requirement — All of the training data must be present in memory in order to calculate the closest K neighbors
* Sensitive to irrelevant features
* Sensitive to the scale of the data since we’re computing the distance to the closest K points



**Artificial neural network**

This concept actually comes from network of neurons. Neurons are interconnected in a complex manner.

**Naïve bayes**

Naïve bayes is a classifier of the text based on probability approach. It is an algorithm based on bayes theorem which is used for finding conditional probability that denotes likelihood of occurrence of given event which has previously occurred. According to bayes theorem of probability equation is

posterior = prior X likelihood

evidence

In the context of news document or article d belongs to class c based on bayes theorem is as below.

p(c|d) = p(d|c) X p(c)

p(d)

Among the given algorithm Naïve bayes algorithm is more preferable over the other algorithms because it provides the high accuracy compared to other algorithm. So we are going apply Naïve bayes theorem for the latest news classifier.

**Step 2:**

In second step we use users’ feedback in latest news classifier. Latest new when arrives in internet that time user can give feedback for that news.

After arrival of news there is specific time 10-20 minutes users are able to give answer one question which type of news like Technology, Sports, Business, Finance, Global, Health, Food, Fashion, Entertainment, Bollywood/ Hollywood so that classify news comes into the which class of news. Within just few

minutes lots of people view that news from the feedback of those news algorithm can classify the news comes under which class. After applying two-way classifier, we get maximum accuracy in classifying latest news.

**3.Conclusion**

As Classification of the latest news is require for the giving the news to the user according to their choice we have to classify the news into different categories. For classify the news according to analysis it is derived that Naïve bayes algorithm is best for latest news classification it has higher accuracy more than 80% and after applying this algorithm we will decide the category of news and also after releasing news for first 10-20 minutes we will ask the user about the news that news belongs which category based on reviews of user as well as result of algorithm we will finalize the category or class of the news. This approach will give us the more accurate result compare to applying only algorithm for classifier.

**4.References**

[1]Dipak Kawade and Kavita Oza “News Classification: A Data Mining Approach” December 2016

<https://www.researchgate.net/publication/312035496_News_Classification_A_Data_Mining_Approach>

[2] Lilima Pradhan, Neha Ayushi Taneja, Charu Dixit ,Monika Suhag “Comparison of Text Classifiers on News Articles” International Research Journal of Engineering and Technology (IRJET) Volume: 04 Issue: 03 | Mar -2017

<https://www.irjet.net/>

[3] Jasneet Kaur,Seema Bhagla “News Classification Using Naïve Baye’s Classifier” International Journal of Advanced Research in Computer Science and Software Engineering Volume 6, Issue 4, April 2016

<http://ijarcsse.com/index.php/ijarcsse>

[4] Gurmeet Kaur, Karan Bajaj “News Classification and Its Techniques: A Review” International Organization Of Scientific Research -Journal of Computer Engineering(IOSR-JCE) Volume 18, Issue 1, Ver. II

<http://www.iosrjournals.org/>

[5] Amritpal Singh, Sunil Kumar Chhillar ,“A Survey on the Machine Learning Techniques for the News Classification” Vol. 8, Issue 2, April - June 2017

<http://www.ijcst.com/>

[6]Article: Top 10 Machine Learning Algorithms

<https://www.dezyre.com/>

[7]Article: Naïve bayes classifier

<https://en.wikipedia.org/wiki/Naive_Bayes_classifier>

[8] support vector maching<https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python>

[9]decision tree

<https://www.datacamp.com/community/tutorials/decision-tree-classification-python>