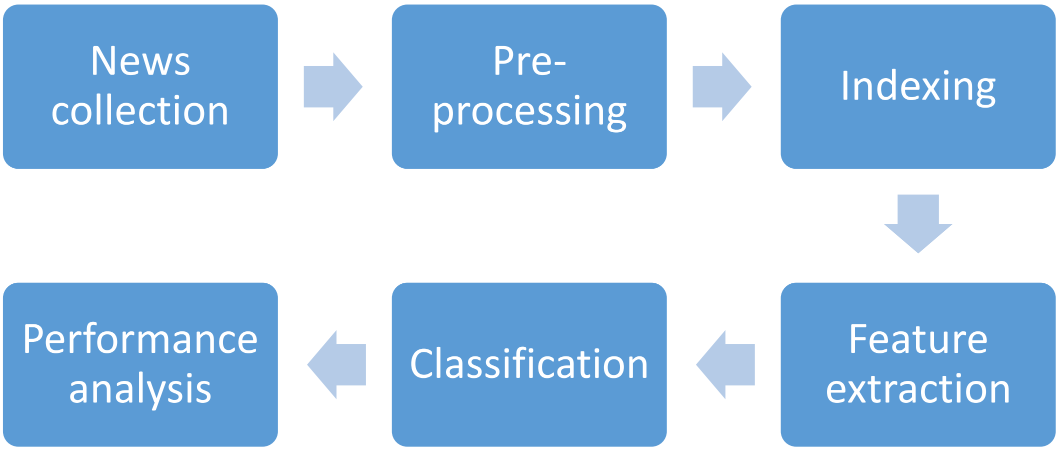
LATEST NEWS CLASSIFIER

Introduction:

In this digital era, huge amount of information is available in digital format. So, to improve machine's decision making, it is essential to classify and analyse such data. With the capability of data mining we can extract useful data from huge databases and can classify the type of news in a given newspaper or article. As we all know, unprocessed data maybe in unstructured format. So, to classify data, it must be in some format. So, before using this data we have to pre-process it, which is quite challenging. The categorization/classification of the news helps users a lot. For ex. A person who is more interested in sports news will get easy access to it and he/she not have to look at the entire newspaper for it. In this type of newspaper, the news is been classified based on their headlines in a particular article or newspaper.

Workflow:



1. **News Collection:**
   The very first step of news classification is news (data) collection. As we are going to train our model on this collected news, we will require huge amount of news data. The news are collected from the various resources like magazines, different types of newspapers, articles, www etc. and also these files can be gathered in any format like .doc,.pdf etc.

2. **News Pre-Processing:**

This collected data may contain some useless, extra information such as tokens, stop words and stemming which must be pre-processed otherwise we may get misclassification. So, to get higher accuracy and correct result data must be pre-processed.

- **News Tokenization**: First of all, obtained data will be in long sentence format. So, in this step, as the name suggest we will fragment each sentence into words (tokens). Now, in the next step we will work these words.

- **Removal of stop words**: This step is used to remove stop words. Stop words means words with no meaning like pronouns, conjunctions, propositions, semi-colon, full-stop, special character etc. These words must be removed as they may result in misclassification. Approximately there are 152 stop words.

- **Word Stemming**: In this step, we will convert every word to its root word by removing suffixes like ing, ed etc. There are different types of stemmers are available like Porter Stemmer, S-stemmer etc.

3. Indexing:

Indexing is the most important step for news classification. Here, the bag of words approach is used to reduce the complexity and difficulty in news classification. Each word in the news content or headline is considered to a vector. Bag of words consist of two things: 1. Vocabulary of known words and 2. Measure of presence of known words and this is used for indexing news headlines. For each word a complete matrix is made as they are in form of vector.

4. Feature Extraction:

When there exist a huge number of features and each of the features is a dominant, primary word for each category, classification process may take long time and accuracy may get affected. To surpass these issues, a process called feature selection is chosen in which only those primary, relevant and highly effective features are chosen, which will give us better news classification.

A large number of techniques exists in literature for selecting relevant features like

- Boolean weighting
- Class Frequency Thresh holding
- Term Frequency Inverse Class Frequency
- Information Gain.

1) Boolean Method Of information retrieval:

This is the most common and primary method used for information retrieval. It uses Boolean logic and set theory concepts in which the data to be searched and query by the user is seen as a set of terms. Queries are considered as a Boolean expression on terms. It uses exact matching scenario for finding the match between the terms to be found in documents. The queries are logically checked by the AND , OR and NOT operators which uses the operators to find the set of documents which contain the "exact words" in the query.

For Instance : x AND y where x and y are elements from the set of terms of query. So the documents which will both x and y will be included in the designated output by Boolean method .

The Boolean model is represented by F, D,Q and R where

F : Boolean algebra over sets of terms and set of documents

D : Set of Indexing terms (keywords) present in the documents that each term is present or not. If the term is present then it is marked as 1 in Set or marked as 0 in Set of indexing terms.

Q : A Boolean Expression which is query needed to be satisfied which consists of related keywords using AND, OR and NOT operators

R : A document is predicted as output which satisfies the query expression.

AND (^) : Intersection of two sets of terms.

OR (v) : Union Of two sets of terms.

NOT (~) : Set inverse

Advantages of Boolean Method :

- Clean and exact formalism
- Easy to implement

Disadvantages of Boolean Method :

- Exact matching (strict matching) leads to too few or too many matching documents.
- As every term is of equal weight , priority classification is not possible.
- As it works on exact matching , more than information retrieval it's more seems like "data retrieval"
- Translation of query in Boolean expression is sometimes difficult.

2) Information Gain :

Information gain measures the probability of particular dataset element occurring to a given value of random variable value. A large value of information gain means that the probability of any data element in this case , Any particular word occurring in the news articles for any particular group. Entropy denotes how much information there is a random variable(keyword) or more precisely it's probability distribution in all the groups.

It is the amount of information gained about a single random variable by observing another variable whose value is known in advanced. But in the case of decision trees which is used here in our case. Information gain means conditional expected value of the univariate probability distribution of one variable to that of another variable.

Information gain is denoted by

$IG(X|Y)=H(X) - H(X|Y)$

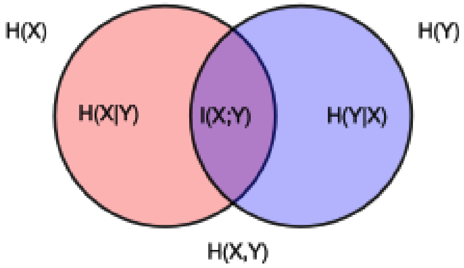
Where,

IG(X|Y) is the mutual information or information gain for X and Y ,

H(X) is the entropy of X

H(X|Y) is the conditional entropy of X given Y

The value of information gain is always greater than zero , higher the value of IG, higher is the relationship between two variables.



Venn representation of Mutual Information

The above figure denotes the additive and subtractive relations among the correlated variables X and Y with their respective entropy.

Uses:

In feature selection and feature transformation in machine learning , used to characterize both the relevance and redundancy of variables

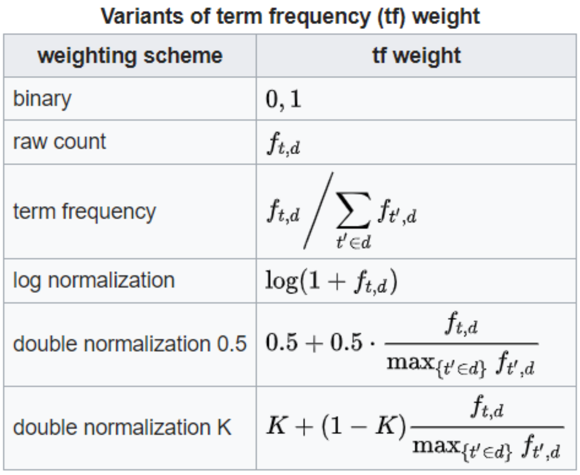
3) **Term frequency inverse class frequency:**

It is denoted by tf-idf , It is used to weight each word. This weight is used to determine the importance of word. Weight(importance) of the word is taken into account during classification. So, weight of the word plays important role in data classification.

The tf-idf value is retrieved by the number of times a word occurs in a document. If the occurrence of word increases, tf-idf also increases. According to a survey held in 2015 , 83% of text-based recommender system uses tf-idf method.

Tf-idf can be used to effectively for stop-words , including text summarization and classification.

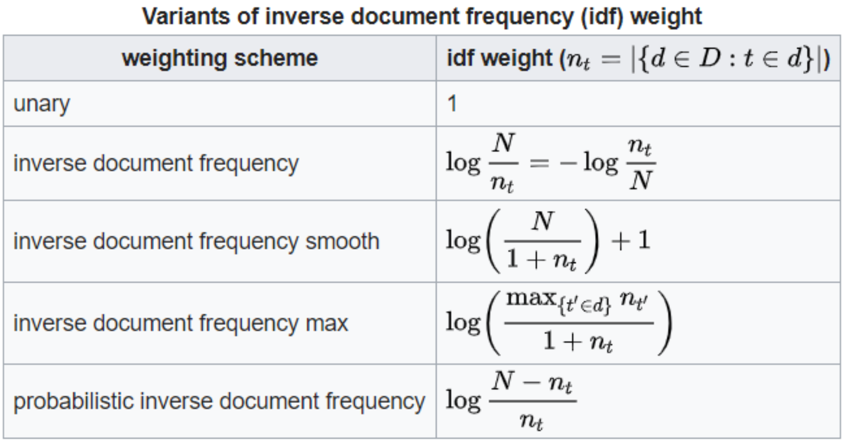
Term frequency: Term frequency is used in data mining. It is defined by the number of times any word has occurred in the document; Weight of a term is directly proportional to the term frequency.



Variants of term frequency (tf) weight

weighting scheme	tf weight
binary	$0, 1$
raw count	$f_{t,d}$
term frequency	$f_{t,d} \Big/ \sum_{t' \in d} f_{t',d}$
log normalization	$\log(1 + f_{t,d})$
double normalization 0.5	$0.5 + 0.5 \cdot \frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$
double normalization K	$K + (1 - K)\frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$

Inverse Document Frequency: IDF is used to decrease the weight of stop words, because stop words occurs excessively in the statement and if we don’t use IDF then the weight of stop words will increase and it will result in wrong classification of data.



Variants of inverse document frequency (idf) weight

weighting scheme	idf weight ($n_t = \|\{d \in D : t \in d\}\|$)
unary	1
inverse document frequency	$\log \frac{N}{n_t} = -\log \frac{n_t}{N}$
inverse document frequency smooth	$\log\left(\frac{N}{1 + n_t}\right) + 1$
inverse document frequency max	$\log\left(\frac{\max_{\{t' \in d\}} n_{t'}}{1 + n_t}\right)$
probabilistic inverse document frequency	$\log \frac{N - n_t}{n_t}$

$\text{TF-ICF}_w = \text{TF}_w * \text{ICF}_w$

Wh TF = Term frequency for a word “w” and ICF is an inverse class frequency for a word “w”.

Advantages :

- Easy to compute
- Can remove stop-words effectively
- More efficient than other methods used for the same purpose
- Easy to compute similarity between 2 documents

Disadvantages :

- Used as lexical level , so cannot capture the semantics need

4) Class frequency thresholding:

It gives all classes which are used for classification at least 1 output and afterwards completing the whole classification after seeing some threshold value, if any particular class doesn't have minimum threshold then it is considered redundant or not applicable, which causes a flaw in it with words which have less frequency but more important classification criteria.

5. Classification:

This phase is called classification phase. It is performed after feature selection phase. The main objective of this step is to classify pre-processed data in given categories. There are many news classification methods are available, common ones are,

- Naive Bayes
- Artificial Neural Networks
- Decision Trees
- Support Vector Machines
- K-Nearest Neighbours

1) Naïve Bayes Algorithm

The naïve bayes algorithm is a statistical classification algorithm that's used for classification task [2]. It uses Bayes theorem for this task. One of the feature of naïve bayes algorithm is, scalability, means it can be used for large dataset also. As the data is labelled, it is the supervised learning algorithm. To classify the customers the various features are used for example, their names, age, birth-date etc.

The classification has to phases: 1. Learning Phase and 2. Evaluation Phase [2].

Learning Phase:

This phase is used to train our model.In order to train the mode, we feed our model with various datasets. This phase is very important as it is the primary requirement for performance measure.

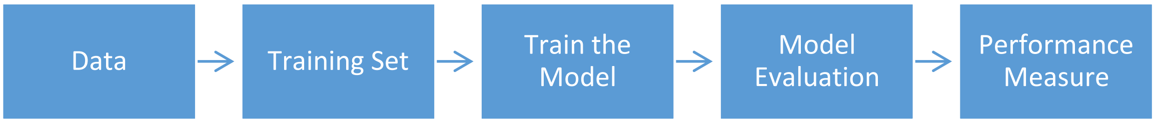


Fig-1 Learning Phase

Evaluation Phase:

In evaluation phase the model is evaluated by giving input to the machine and check whether it works properly or not. Performance of model is measured by model's accuracy, precision and error etc.

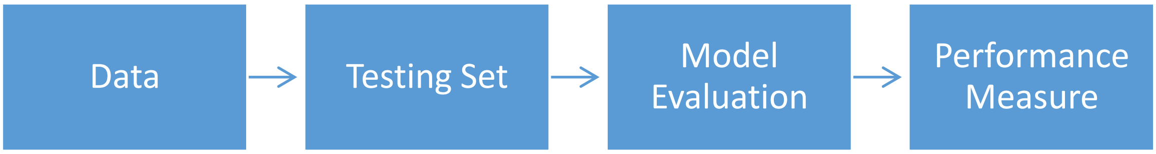


Fig-2 Evaluation Phase

The naïve bayes theorem is based on class conditional independence. Conditional Independence means each and every attribute is independent from each other. This theorem is based on finding the posterior probability, P(class|attribute) from P(class), P(attribute) and P(attribute|class).

Therefore, **Posterior Probability (P(c|a)) = P(a|c) * P(c) / P(a)**

P(c|a) = Posterior Probability of the target class for given attribute

P(a|c) = Probability of the attribute for given class

P(c) = Prior Probability of the target class

P(a) = Prior Probability of the attribute.

To find the Posterior Probability the following steps are followed:

Step-1: Find the probability for each class labels

Step-2: Find the probability of each attribute for each class

Step-3: Put the probabilities in the bayes probability for finding the posterior probability.

Step-4: Observe which class is having the highest probability.

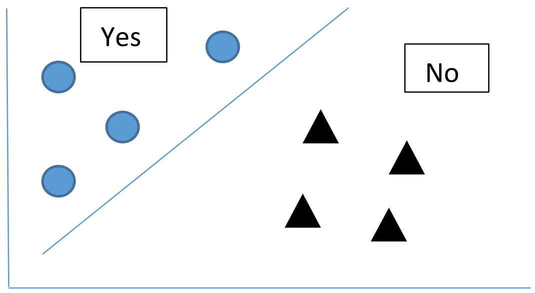
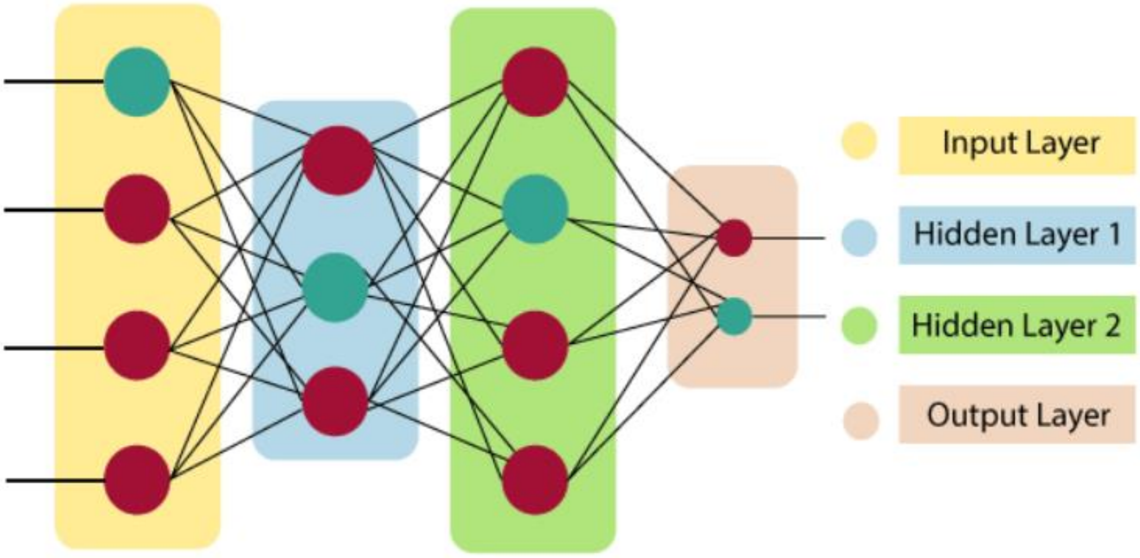


Fig-3 Classification for label yes and no

As It is the classification algorithm, it is used in text classification, sentimental analysis, spam filtering etc.

2) Artificial Neural Networks

Artificial Neural networks is same as neural network of our brain. It is composed of nodes and layers. Node of ANN is same as the neuron of brain. It is the combination of summation of input and weight associated to it and activation function. This node takes input and generates output using activation functions which is same as the action we perform after our neurons sense something.

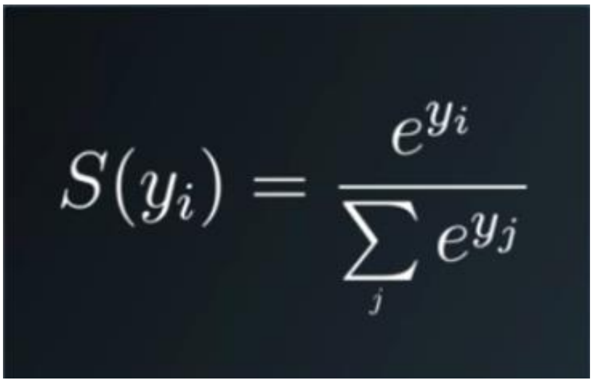


Softmax function:

It is an activation function that takes logits as an input and transforms them into probability distribution. Basically, it produces probability from logits. It is probability distribution because the sum of distribution is always 1.

Logits are the real numbers from -∞ to +∞ which is basically the output of logit layer and the last layer of neural network for classification task is called *logit layer*.

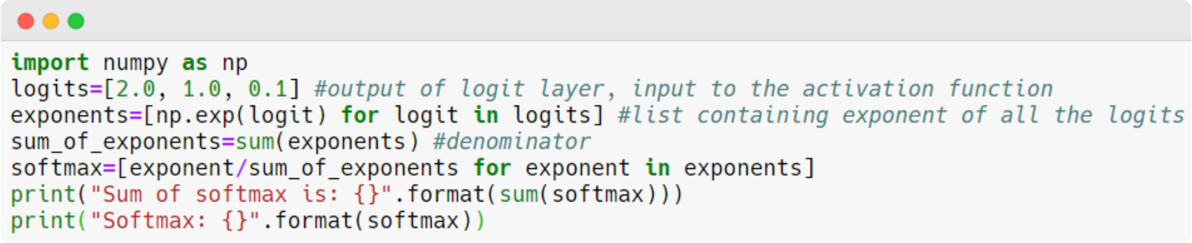
The equation for Softmax activation function is:



$$S(y_i) = \frac{e^{y_i}}{\sum_j e^{y_j}}$$

Where y_i is the input logit number, Σe^{y}_{i} is the sum of exponents of each logit number And it describes Softmax of logit number y_i.

Now we will understand by example how Softmax works. Please refer the below visual.



```
import numpy as np
logits=[2.0, 1.0, 0.1] #output of logit layer, input to the activation function
exponents=[np.exp(logit) for logit in logits] #list containing exponent of all the logits
sum_of_exponents=sum(exponents) #denominator
softmax=[exponent/sum_of_exponents for exponent in exponents]
print("Sum of softmax is: {}".format(sum(softmax)))
print("Softmax: {}".format(softmax))
```

First of all, we have taken the list of logits and generated list of exponent of all the logits means we are done with numerator of Softmax function. Now as we know that, for the final output we require normalized result. So, to normalize it, we will divide each exponent with the sum of exponents. So, our Softmax is ready.

Now, you may have questions like why we have done exponent(logit)/sumofexponents(logits) and why not logit/sumof(logit)?

It is because logits are the real number ranging from -∞ to +∞, So when the logits are negative, sum of them doesn't gives the correct normalization whereas exponents of logits turn them into 0 to +∞.

It is mostly used in classification. Softmax generates large output if the input set is large and vice versa. Generally, it pushes result, close to 0 or 1. If you have decided to use Softmax function as your final classification function then it is advisable to use Cross Entropy loss function.

Behavior:

As the output of Softmax function gives probabilities, the output ranges from 0 to 1. And the sum of outputs is 1.

3) Decision Tree

Decision tree is the supervised machine learning algorithm used for data mining. It is used for both classification task as well as regression task. The data is continuously separated according to certain parameters. Decision tree consists of the common tree components such as Nodes, Edges/Branch, Leaf. Nodes are the testing attribute. The value of the node is passed to the next node or the leaf as an edge/branch. The leaf node predicts the outcome which represents the class labels or the class distribution.

Types of Decision Tree:

- **Classification trees**
- **Regression Trees**

Classification Trees:

This type of tree is used for classification task. When decision tree has categorical variables it is called classification tree. Classification tree are of Yes/No types. Means at each we have two options for the result. At the end the decision label is either Yes or No. Classification tree is built by recursive process of partitioning the data.

For example, the student has done the homework or not is of Yes/No types. Another example is that if the student scored above 70 marks in an exam, this is also a Yes/No types decision.

Regression Trees:

Decision tree becomes regression tree when tree has continuous target variable. The target variable or the end of the decision label can be any real number or any continuous values are called regression tree.

For example, Number of students scored below 40 marks in an exam. Another example is of how many students have not done the homework. This takes any continuous value or a real number.

The most important feature of decision tree is it doesn’t require normalization of data. Another benefit of decision tree that is, it is easy to understand. It has capability of decision-making knowledge regardless of the length of the supplied data. Missing values are handled also. The rules generated from the decision tree are mutually exclusive which means that no two rules conflict and exhaustive which means that there is only one attribute-value combination for each rule. So, the order of the rules does not affect the final result.

Decisions are stored at the leaf node of decision-tree.

4) Support Vector Machine (SVM)

Support Vector Machine algorithm is also known as SVM. It is the best supervised learning algorithm that can be used for both classification task as well as regression type of task. It has good accuracy and precision in classification task as well as regression task [1]. It can be used for linear as well as non-linear problems also. This algorithm is used to find a line that distinctly classify the given points perfectly in case of classification. There are many separating hyperplanes possible to classify the samples, but we have to choose that hyperplane that maximizes the margin [1]. Support vectors are the data points that are close to the hyperplane (minimum perpendicular distance) & which has impact on the shape and position of hyper plane. As it impacts the shape of hyperplane, removal of any those data points, could change the shape and position of the hyperplane [1].

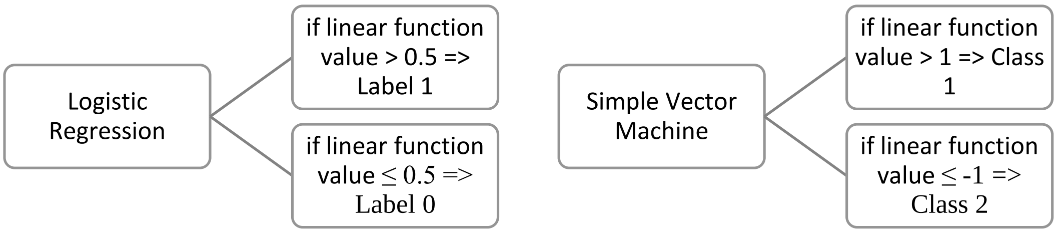

Logistic Regression
if linear function value > 0.5 => Label 1
if linear function value ≤ 0.5 => Label 0
Simple Vector Machine
if linear function value > 1 => Class 1
if linear function value ≤ -1 => Class 2

Fig 1. Comparison between linear function in case of Logistic regression & SVM

In SVM, our task is to find maximum margin between data points by creating a line between points of two different classes [1]. To maximize the margin between data points, hinge loss function is used.

$$H(x,y,f(x)) = \begin{cases} 0 & ;\ when\ y * f(x) \geq 1 \\ y * f(x) & ;\ else \end{cases}$$

For the given sample set containing two types of samples: red-cross & blue-triangle as shown in Fig 2. & 3., we are supposed to classify the sample into two groups. There are two feature vectors and a graph are plot for both the samples. Now in Fig 2., we can observe that there are two clusters formed one of red-cross on left side and another of blue-triangles on right side. Here the optimal hyperplane cloud be straight line. Thus, this type of hyperplane is known as linear hyperplane.

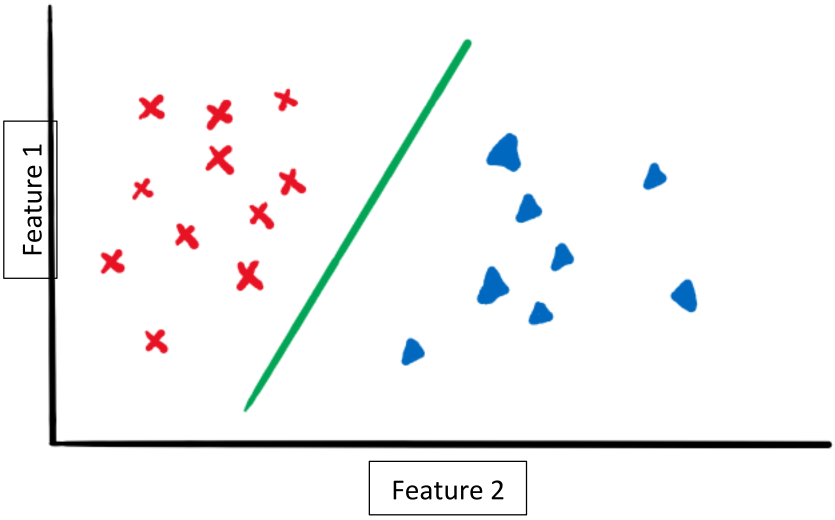


Fig 2. Linear Hyperplane

For the data points which are scattered in the plane in non-linear fashion like one mentioned in Fig 3. could be classified using a non-linear hyperplane like circle, parabola, hyperbola, etc.

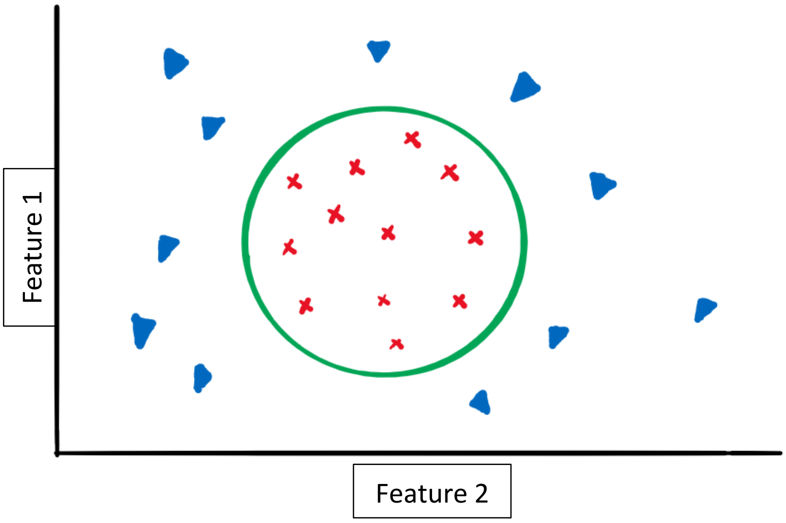


Fig 3. Non-Linear Hyperplane

5) K-Nearest Neighbor (KNN)

K nearest neighbor also known as KNN. It is one of the simple supervised learning algorithms that saves all available cases and then classifies the new case based on a measure known as distance function shown in Fig 4.

Euclidean Distance:	$D(X,Y) = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$
Manhattan Distance:	$D(X,Y) = \sum_{i=1}^{k} \mid x_i - y_i$
Minkowski Distance:	

$$D(X, Y) = \sqrt[q]{\sum_{i=1}^{k} (|\, x_i - y_i \,|)^q}$$

Fig 4. Various distance functions used in KNN algorithm.

In the above equations D(X,Y) is the distance between the new case X and the previously stored cases Y, k is the number of neighbors considered and can be any integer.

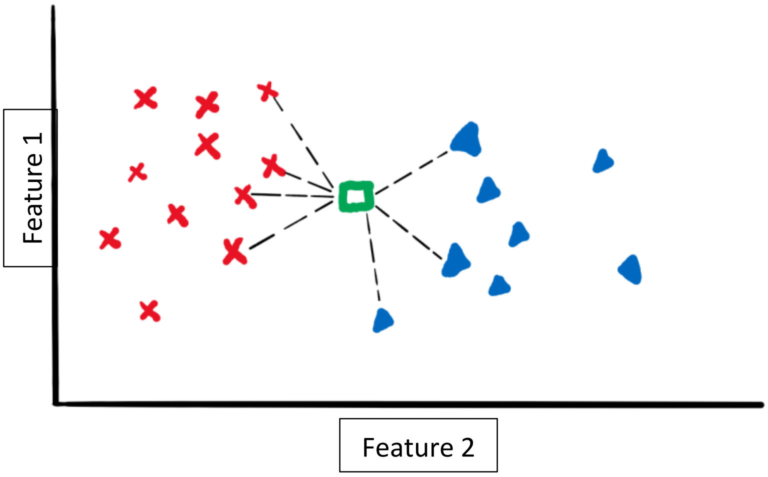


Fig 5. KNN plot

Generally, number of neighbors k is determined by parameters tuning. D(X,Y) is applied with all the previous cases Y and X is added to the cluster which yield minimum distance value of all the cases calculated.

KNN algorithm consists of the following steps:

1. Decide and Initialize number of neighbors be considered k.
2. Calculate the distance.
3. Find the closet distance.
4. Vote for labels.

6. Accuracy, Precision, Recall, F1 Score

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Accuracy - Accuracy is the foremost intuitive performance live and it's just a relation of properly predicted observation to the general observations. One may think that, if we've high accuracy then our model is best. Yes, accuracy is also pleasant live but solely you've got symmetric datasets where values of false positive and false negatives unit of measurement nearly same. Therefore, you've got need to appear at different parameters to gauge the performance of your model.
Accuracy = TP+TN/TP+FP+FN+TN

Precision – Precision is that the relation of properly predicted positive observations to the total predicted positive observations. High precision relates to the low false positive rate.
Precision=TP/TP+FP

Recall (Sensitivity) - Recall is that the relation of properly predicted positive observations to the all observations in actual class -yes.
Recall = TP/TP+FN

F1 score - F1 Score is the weighted average of precision and Recall. Therefore, this score takes every false positives and false

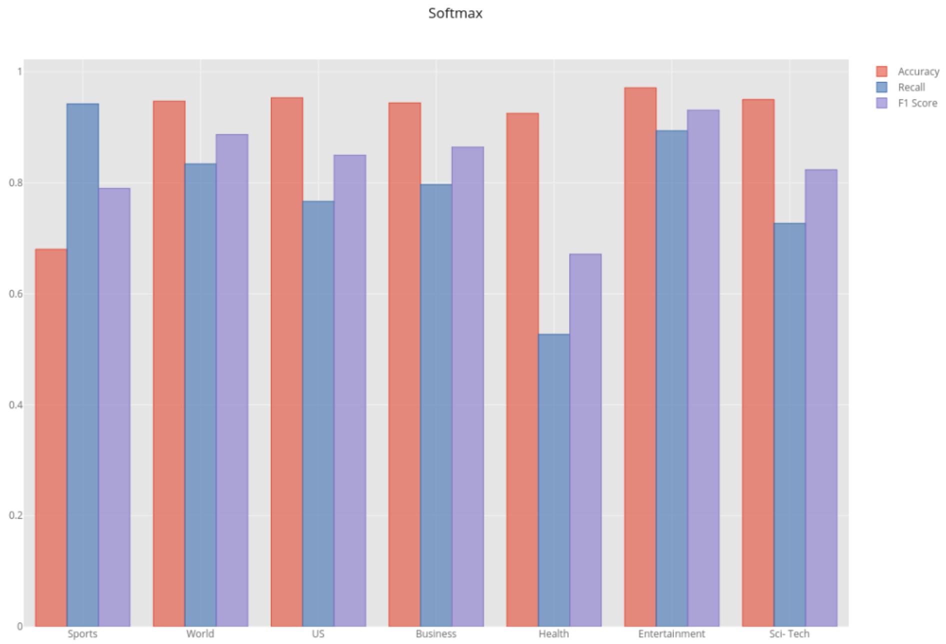
negatives into consideration. Intuitively it is not as simple to understand as accuracy, but F1 is typically further useful than accuracy, significantly if you've got an uneven class distribution. Accuracy works best if false positives and false negatives have similar value. If the worth of false positives and false negatives are totally different, it’s better to appear at every precision and Recall.
F1 Score = 2*(Recall * Precision) / (Recall + Precision)

True Positive Rate(TPR)= TPR is the ratio of True positive i.e. the predicted positives which are truly positive to the Total positive value.
TPR = TP / P = TP / TP+FN

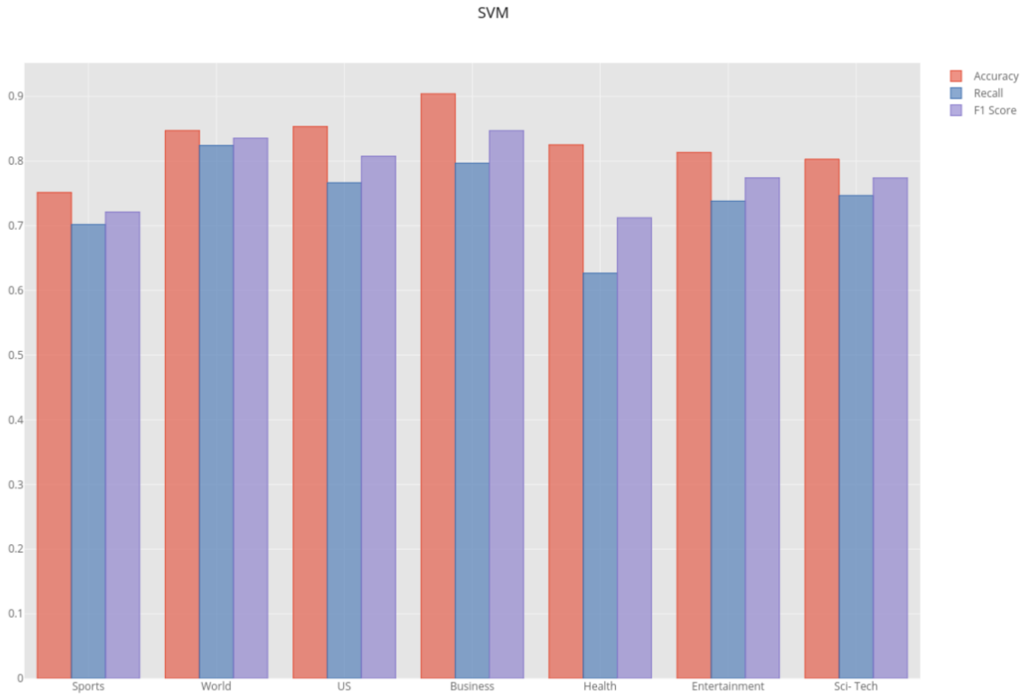
False Positive Rate(FPR)= FPR is the ratio of False positive i.e. the predicted positives but are truly negative to the Total negative value.
FPR = FP / N = FP / FP + TN

7. Graphs

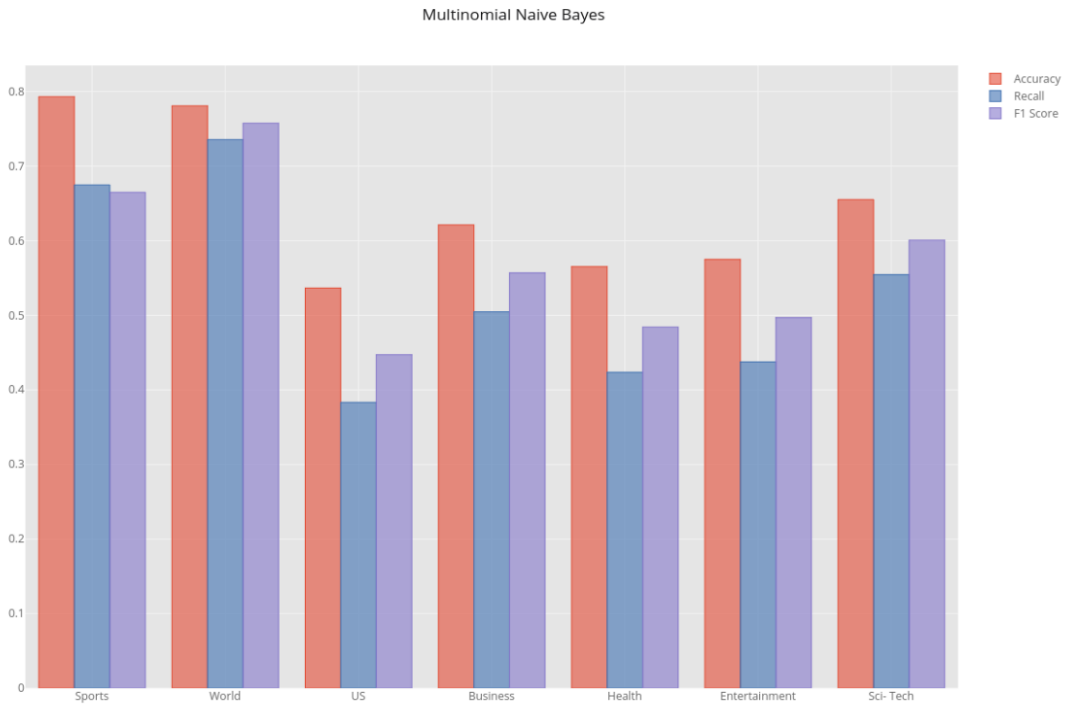
Softmax



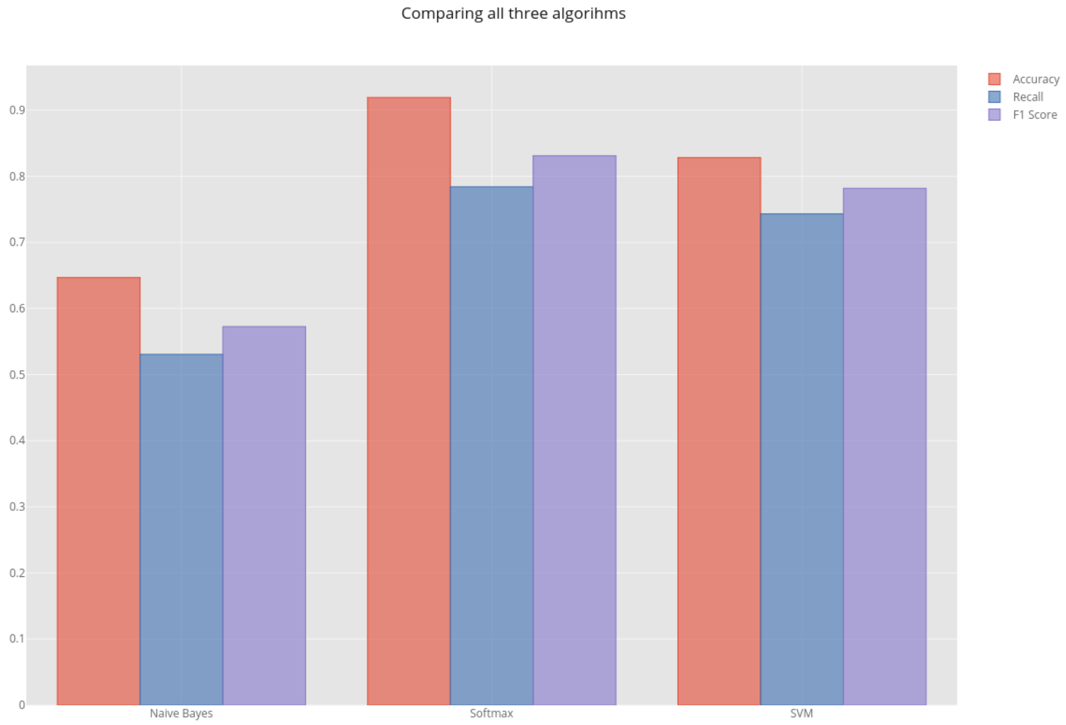
SVM



Multinomial Naïve Bayes



Comparison of All three algorithms



- **Conclusion:**

  We have seen different algorithms used for news classification, from these algorithms we plotted graphs for 3 algorithms that are Softmax which is part of MLPClassifier, SVM, and Multinomial Naïve Bayes. We used the graphs for comparison between these 3 algorithms and the last graph gives comparison of all three algorithms from this we can interpret that Softmax has the best Accuracy, Recall and F1 Score followed by SVM and Naïve Bayes with the least scores.

8. References

[1] Rohith Gandhi 2018, *Support Vector Machine — Introduction to Machine Learning Algorithms*, Towards Data Science, accessed on 8 March 2020, <https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47>.

[2] https://www.datacamp.com/community/tutorials/naive-bayes-scikit-learn