LATEST NEWS CLASSIFIER

ABSTRACT

Throughout the previous years, content mining has an increasing noteworthy significance. Since Knowledge is presently accessible to clients through an assortment of sources for example electronic media, advanced media, print media, and some more. Because of gigantic accessibility of content in various structures, a ton of unstructured information has been recorded by investigating specialists and have discovered various courses in writing to change over this dispersed content into characterized organized volume, normally known as content order. Concentrating on full content arrangement for example full news, enormous archives, long length writings and so forth is progressively unmistakable when contrasted with the short length content. In this project, we talked about content grouping procedure, classifiers, and various element extraction systems however all in the setting of short messages for example news characterization dependent on their features.

INTRODUCTION

In this era of digitized data, there has been a constant need for accurate data classification and getting valuable information from the varied forms of data available. Text Mining techniques play a very significant role in extraction of valuable information. Text mining is the finding of some information which is previously unknown by extracting that information from large sets of unstructured text. Such type of unstructured information or text cannot be used for any further computer processing or programming. Hence, we require some appropriate processing and preprocessing on the text data to extract valuable and meaningful information from this unstructured data. There is much research going on in the field of text mining as there is necessity of text classification in almost all the fields, whether it be medical or any technology related field. The major research work is going on the news classification. Our main focus in this research is to carry out news classification based on their headlines. A huge variety of news headlines classifications have taken place in the past few years, few of them include emotions classification on basis of news headlines, financial news classification, classification of short texts, automatic news headlines classification, news headlines classification using N-gram model, classification of news headlines for providing user centered e-newspaper, emotions extraction from news headlines, and short news headlines classification of twitter and many more. But in this paper, we have decided to classify the news based on headlines into 4 major categories. The categories are as follows:

1. Sport
2. World
3. US
4. Business
5. Health

6. Entertainment
7. Sci-tech

We have performed a comparative study between 3 different classification approaches , which are Naive Bayes, Softmax and SVM(Support Vector Machine) algorithms.

LITERATURE REVIEW

Majority of the information is stored in the form of text like emails, newspaper articles, research reports, letters from customers and company reports. In case of Newspapers, news is provided under various categories like International, National, Politics, Sports, Entertainment, Finance, etc. Classification of text is an essential part of data mining. Classification begins with the training of a set of documents that are already labeled with class. Text classification consists of two types, namely single label and multi label. A single label file belongs to only a single single. While, a multi label file belongs to multiple classes.

Research in Text mining field is becoming more necessary as a large number of text is now readily available in every field for classification. Several fields such as finance, medical, image processing, etc,. Have a major objective of text mining and to extract useful information by using data mining algorithms.
Data mining and Machine learning Methods are used in a combination to classify and find patterns from various data text from documents, images, etc. The main focus of this undertaking is to carry out a review on the news headlines classification. Full news text consumes a huge amount of time and calculations, furthermore increasing chances of misclassification of the news.

Numerous researchers classified the data by news headlines, which helps in classifying each news headline into its pre-defined class. It works by finding the most suitable words within class and are then selected.Text categorization is the problem of classifying text documents into a set of predefined classes.

The applications for Naive Bayes Algorithm are:

- Real-time Prediction: As Naive Bayes is super-fast; it can be used for making predictions in real time.

- Multi-class Prediction: This algorithm can predict the posterior probability of multiple classes of the target variable.

- Text classification/ Spam Filtering/ Sentiment Analysis: Naive Bayes classifiers are mostly used in text classification (due to their better results in multi-class problems and independence rule) have a higher success rate as compared to other algorithms. As a result, it is widely used in Spam filtering (identify spam email) and Sentiment

Analysis (in social media analysis, to identify positive and negative customer sentiments)

- Recommendation System: Naive Bayes Classifier along with algorithms like Collaborative Filtering makes a Recommendation System that uses machine learning and data mining techniques to filter unseen information and predict whether a user would like.

The advantage of Naive Bayes is that it works with both textual and numeric data and it is very easy to implement and to compute.But the disadvantage of it is it gives bad performance when features are correlated like short texts or news headlines classification. In real world data, conditional independence assumption is poorly violated.

A Support Vector Machine is a supervised learning algorithm that sorts data into two categories. The main task of SVM is to identify which category a new data point belongs to. Thus, making SVM a non-binary linear classifier.

Some common applications of SVM are-

- Face detection – SVM classify parts of the image as a face and non-face and create a square boundary around the face.
- Text and hypertext categorization – SVMs allow Text and hypertext categorization for both inductive and transudative models. They use training data to classify documents into different categories. It categorizes on the basis of the score generated and then compares with the threshold value.
- Classification of images – Use of SVMs provides better search accuracy for image classification. It provides better accuracy in comparison to the traditional query-based searching techniques.
- Bioinformatics – It includes protein classification and cancer classification. We use SVM for identifying the classification of genes, patients on the basis of genes and other biological problems.
- Protein fold and remote homology detection – Apply SVM algorithms for protein remote homology detection.
- Handwriting recognition – We use SVMs to recognize handwritten characters used widely.
- Generalized predictive control (GPC) – Use SVM based GPC to control chaotic dynamics with useful parameters.

The Softmax regression is a form of logistic regression that normalizes an input value into a vector of values that follows a probability distribution whose total sums up to 1. Due to this Softmax is referred to as multinomial logistic regression.

Applications of Softmax are:-

- The softmax function is used in all sorts of multiclass classification techniques, such as artificial neural networks, linear discriminant analysis,e etc.
- It is highly used in the final step of a neural network-based classifier.
- It is also used in the field of reinforcement learning where it converts values into action probabilities.

In this paper we have discussed an intelligent system which designs keywords from newspaper headline dataset and also classify it according to predefined categories. We have used Naive Bayes, SVM and Softmax and also did a comparative study. In this undertaking we have focused on categories : sports, world, us, entertainment, sci-tech, health and business.

METHODOLOGY

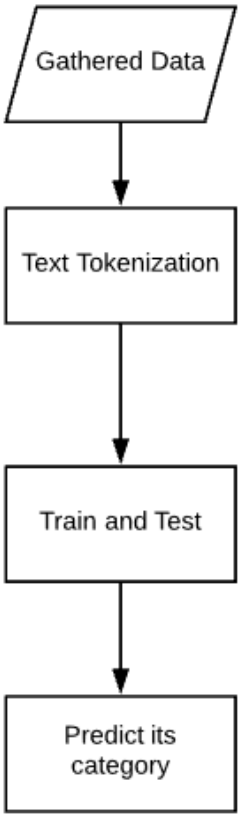
News classification procedure:-

1. We have used the TagMyNews dataset provided by the Computer Science Department of University of Pisa. This dataset contains 32000 English news articles and the categories available are Sport, Business, U.S., Health, Sci&Tech, World and Entertainment.

2. The data has title, description, links and more details and category which is in the text format. Further we open the news file and extract the data in the Data folder from which we get the snippets of each category in the respective folder named as per the categories. We have looped through all the datasets and saved it in a training data list using glob library. We then saved it as a pandas data frame. Moreover, we split the dataset in train and test dataset. We have run the predict function on the test dataset and print the output to later compare with the actual category. Again we have used MLPClassifier and split the data into test and train and further predicted the output and compared it with the actual one. Furthermore, we have done the same for the SVM algorithm.

3. We have used CountVectorizer to get the count of each type of word which is there in the dataset. In TF-IDF it counts the frequency of a word in each document and the more it appears across all the documents the less important it becomes which is completely opposite if it occurs more in the same document.

4. The result was generated for all the three algorithms to calculate the accuracy of prediction for each algorithm. Comparing all three algorithms we got highest accuracy in the softmax algorithm above 90%.



RESULTS

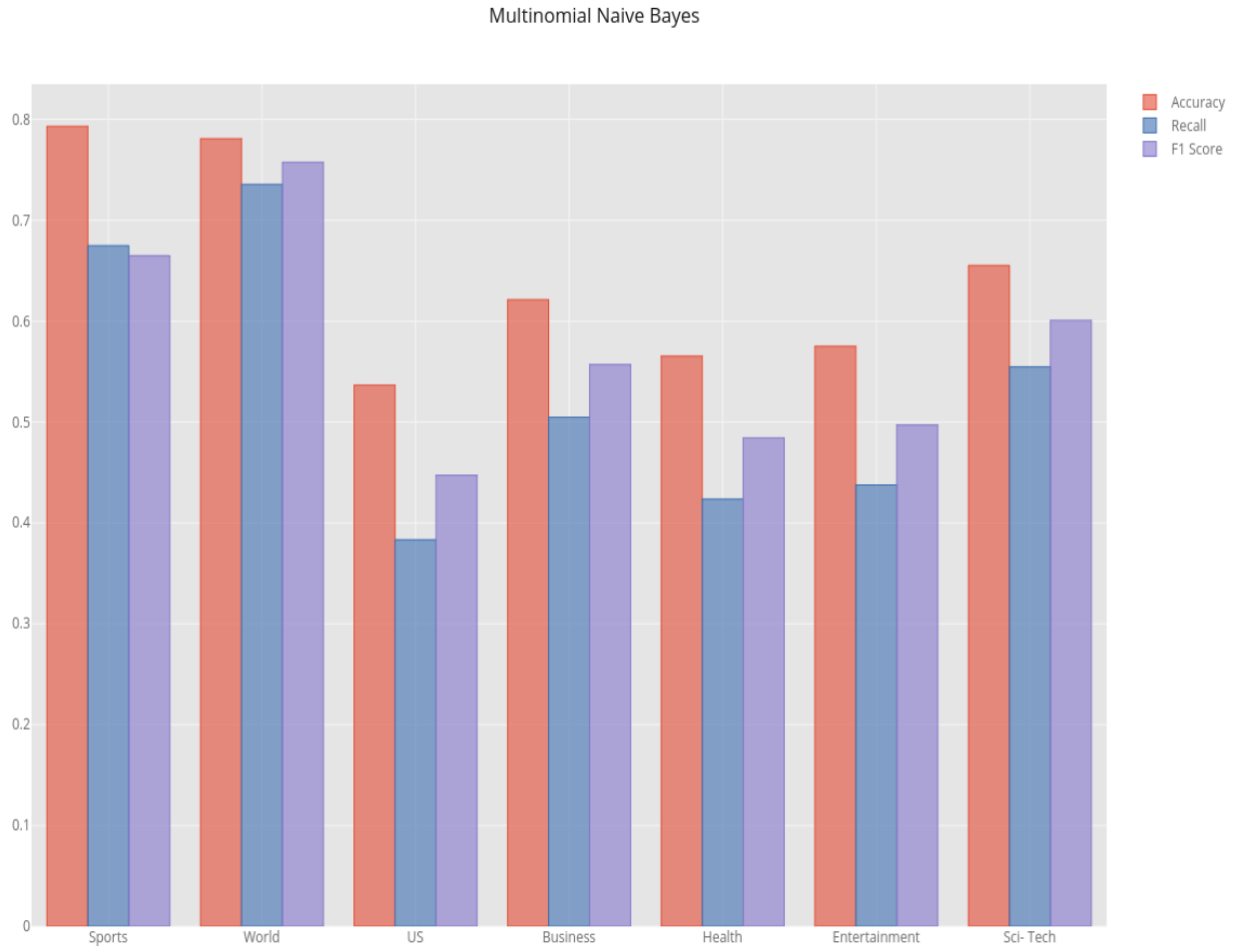
We used Naive Bayes, Support Vector Machine(SVM) and Softmax for the classification. We defined the algorithm on three parameters which are Accuracy, Recall and F1 scores.

Result from Multinomial Naive Bayes:

We used the TF-IDF approach

Avg. Accuracy : 64.3%

Here, We get good accuracy for only a few categories while for the rest of categories we get a 50% accuracy. Moreover Recall and F1 Score average is also quite similar to the accuracy.

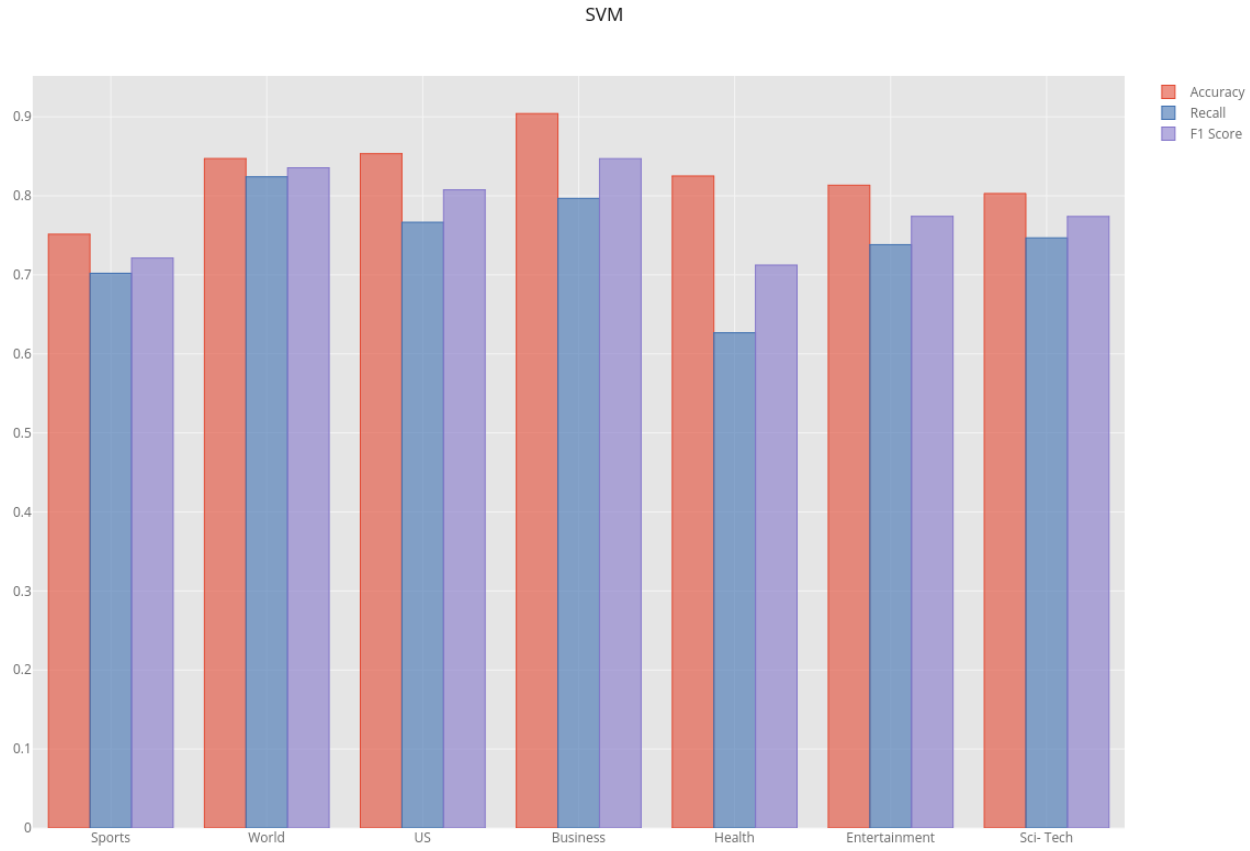


Result from Support Vector Machine(SVM):

We used the TF-IDF approach.

Avg. Accuracy : 82.5%

Here, we get an average accuracy of 82.5% where only the business category has a higher accuracy of almost 90% which is somewhat more than other categories. The average of recall and F1 score values are 85.2% and 78.1% respectively.

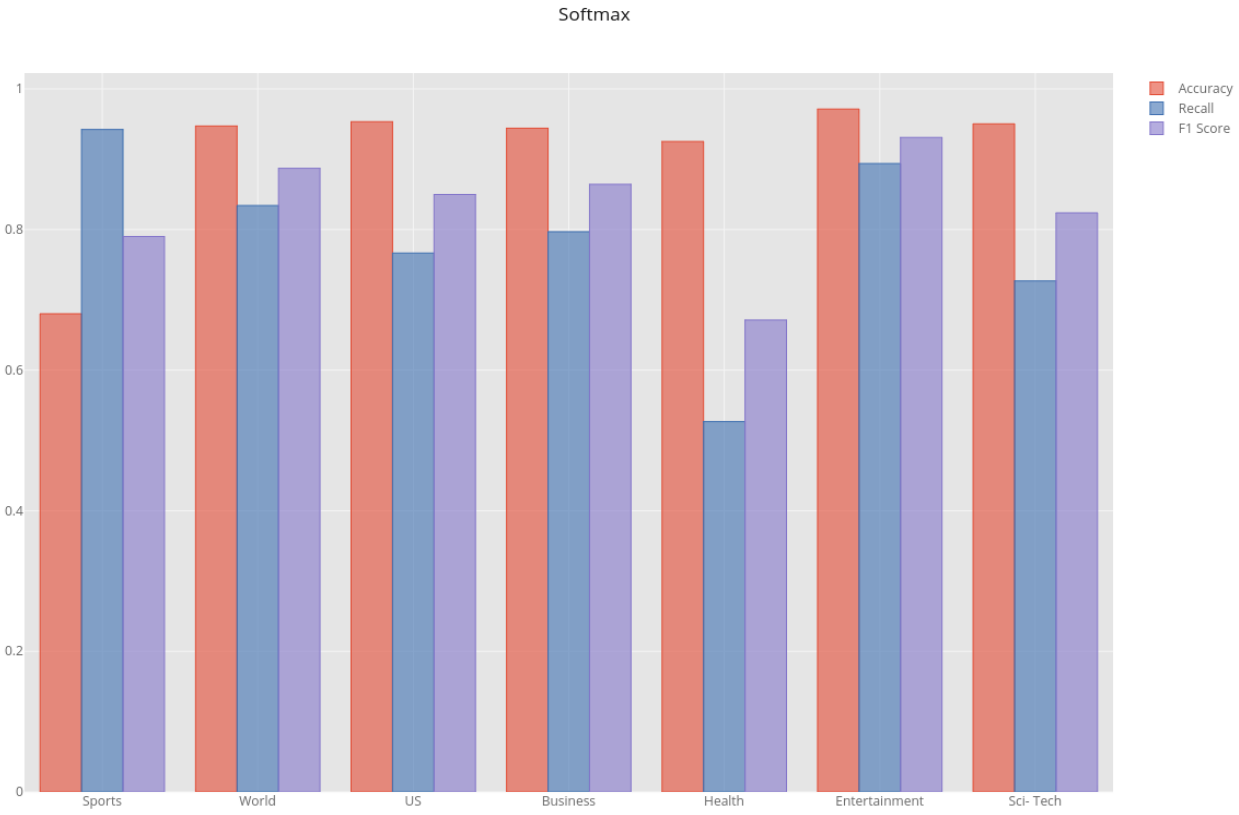


Results from Softmax:

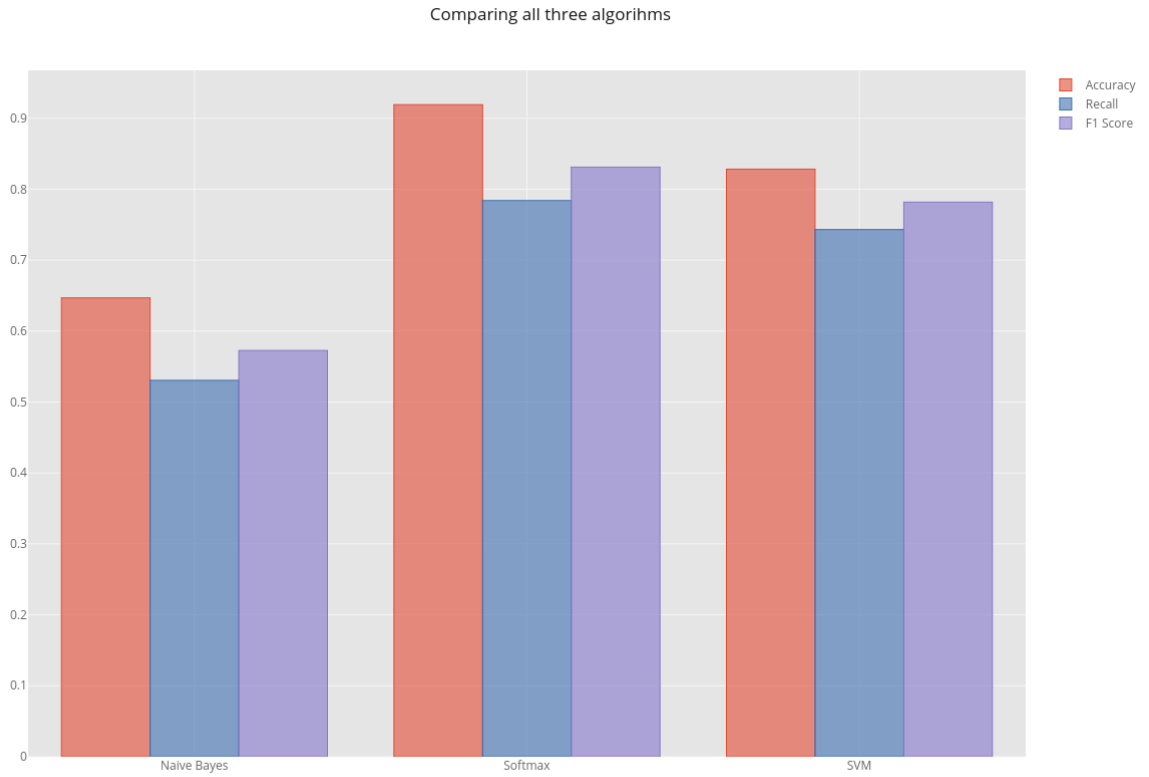
We used the TF-IDF approach.

Avg. Accuracy: 90.4%

Here, the average accuracy is 90.4%. We only get less accuracy in the sports category while in the majority category we get >90% accuracy. The average recall and F1 score are 77.4% and 82.9% respectively.



Thus, overall it can be seen by studying the overview of all the graphs that Softmax is the most effective for classifying news headlines from a variety of categories of a newspaper with an average accuracy of 90.4%. Secondly, Support Vector Machine(SVM) also shows pretty good average accuracy with 82.5%. On the contrary, Naive Bayes shows the most poor performance for classifying news headlines with an average accuracy of just 64.3%. Moreover, SVM and Softmax only show poor accuracy for only 1-2 categories while Naive Bayes classifier shows higher accuracy for only 1-2 categories



DISCUSSION

The areas of text mining is so vast that anything is possible to improve the systems accuracy results. As per the reports the process workflow for the classification remains constant. Few of the changes were recorded in feature selection and pre-processing. Most importantly, accuracy factor and statistical calculations have been reduced in news headlines classifications, which eventually reduce the time as well as increase overall system accuracy. Each algorithm is important at its own place but the right choice of classifier must be made after taking a look at its advantages as well as disadvantages listed with it.

CONCLUSION

In this paper, a review of news headlines classification is done. All the steps i.e., data gathering, text tokenizing, training, testing and predicting of the dataset are explained in detail by the literature beforehand. Comparative study between different algorithms and their accuracy for the given dataset is also discussed at the end. From the above discussion it is observed that only Softmax classifier can be considered as an appropriate classification technique for News headlines classification approach. Different classification scenarios and algorithms perform differently depending on news and data gathered.

REFERENCES

1. https://www.datacamp.com/community/tutorials/naive-bayes-scikit-learn
2. http://cs229.stanford.edu/proj2018/report/183.pdf
3. https://www.kaggle.com/kinguistics/classifying-news-headlines-with-scikit-learn
4. https://towardsdatascience.com/support-vector-machineintroduction-to-machine-learning-algorithms-934a444fca47.
5. https://www.pyimagesearch.com/2016/09/12/softmax-classifiers-explained/
6. https://medium.com/datadriveninvestor/notes-on-deep-learning-softmax-classifier-971b3df27466
7. News Classification Based on Their Headlines: A Review by Department of Computer Engineering, Bahria University Islamabad.