**News Category Classification from News Description**

**Project Report**

**CMP7161 Advanced Data Science**

Date: 10 May 2021

**Abstract**

Over the last few years the number of news articles published on various websites has increased drastically. And newspaper articles are not limited to a few topics. Newspaper covers a huge range of interests from sports to politics to entertainments. So manually tagging this wide variety of news category is more complex as well as it takes huge time. For this now we have started using some technology to automatically classify news category based on news article content.

In this projects we evaluates **Naive Bayes and Decision Tree** machine learning algorithms. And tried to automatically classify news category based on news content. To experiment the system, a dataset from BBC news[1] is used here.

**1. Introduction**

Due to the speedy growth of digital systems, the text data amount is increasing very fast. When we keep together those data it has no meaning. These are unclassified data. With all this, often unstructured, online information available, there is an increasing need for techniques that give structure to this information. Classifying news category is one such problem. In the 80’s, when constructing an automatic classification system this was done by building an expert system. The system consists of simple ‘if-then-else’ rules, which are used to classify a piece of data into a category. In other words, the domain knowledge of experts is translated to rules which are easily interpretable by a computer system. Although this approach can achieve good performance, a clear drawback is that the rules have to be constructed manually. Usually this is done by people other than the domain-experts, which have explicit knowledge about constructing those rules. This makes it difficult to expand to a new domain or to add new categories. The machine learning approach to automatic text classification solves this problem. The only thing needed is an example database with pre-classified examples. Instead of manually creating rules, now a model is created automatically by learning from examples.

In this project different Machine Learning and Natural Language Processing techniques are used to predict which genre or category a piece of news will fall in to from the story.

**2. Background**

To classify test data, various machine learning techniques like SVM and Random Forest have already been used in different research. Using all those algorithms first we train our model and further for new data we try to predict output using our trained model. In this project, we also build some ML-based classifiers to classify news categories based on the next content. In the past few works have been done for news category classification using news headlines and financial news classification [2]. Work has been done on automatic news headlines classification [3]. Few of them built news headlines classifier using N-gram techniques [4]. Others have worked on extracting emotions from news headlines [6]. After taking the data from Twitter built short news headlines classification [7]. In some worked [8] they used TFIDF and SVM classification techniques. They used data pre-processing, then feature selection based on TF-IDF. They grouped by using an SVM classification techniques. On same work they have used BBC news and 20newsgroup dataset.

**3. Aim & Objectives**

The objective of this project is to build a smart news category classifier by using some machine learning algorithms. And predict actual category for new news content.

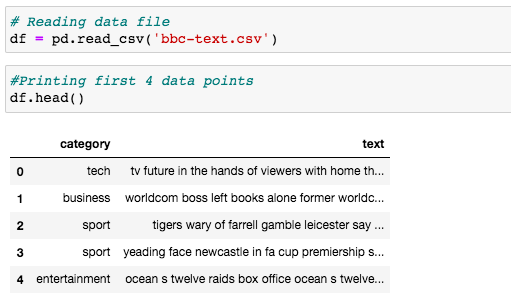
**4. Problem Statement**

News classification is the task of classifying news categories based on news text content. And to achieve that we can use state of art machine learning techniques.

Lots of work on classify news categories has been done in the past. Some of them have built automatic news headlines classification [3] using SVM, Decision tree. Others have worked on news headlines categorization using the N-gram model [4]. Here [8] they also worked on automatic news category classification problem using Tf-IDF and SVM techniques.

**5. Dataset:**

To experiment the system we have used here a dataset from BBC news [1]. There are 2225 documents in this data set. Data are from the BBC news website. And there are five class labels in this dataset. These are business, entertainment, politics, sport and tech.

****

**6. Proposed scheme**

**Summary:** Our goal in this project is to build a classifier that can determine what genre of news a given article or a forum post related to. Ideally, such a classifier would both be able to identify most of the categories widely used in the world of news.

In this project, we evaluate some machine learning techniques. Here we have used Naive Bayes and Decision Tree for news category classification problems. For this project work we have used a dataset from BBC news [1].

**Workflow:** Here is the entire work flow to train machine learning model.

Import data

EDA

&

Data Visualization

Result

Prediction

ML Model

Train data

Test data

Tokenization &

Feature Selection

Train Test Split

Data Pre-processing

**7. Exploratory Data Analysis:**

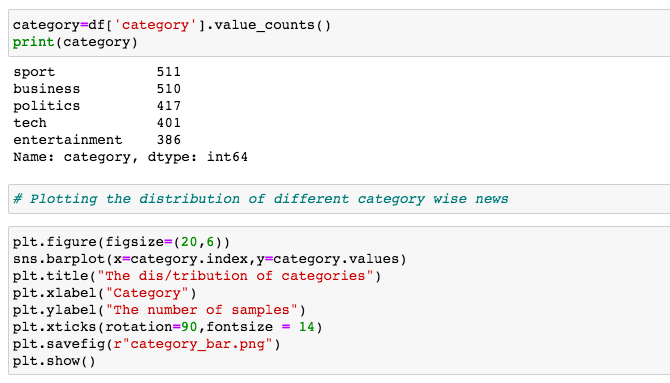
We use Exploratory Data Analysis to summarize their basic characteristics. It is use to draw different graphics and other data visualization techniques. A good EDA gives us a brief idea about different data distribution. We generally do EDA before applying any machine learning model.

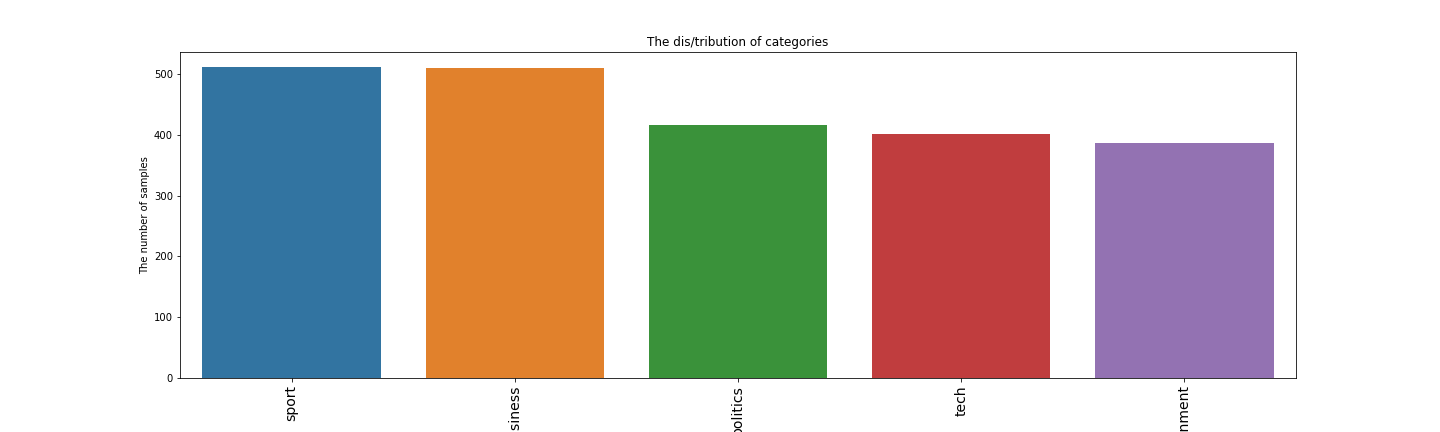
**Different news category analysis:**

Below table shows the category wise number of data points.

|  |  |
| --- | --- |
| Category | Number |
| Sport | 511 |
| Business | 510 |
| Politics | 417 |
| Tech | 401 |
| Entertainment | 386 |

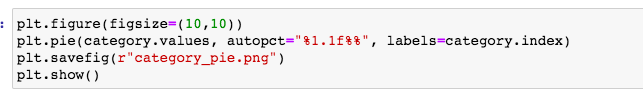
Now, plotting the distribution of different news category:

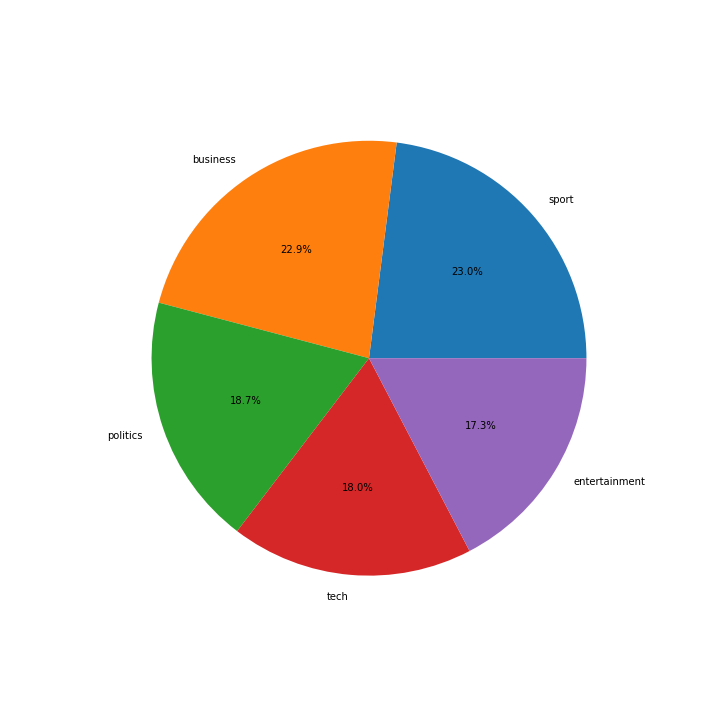




The distribution of category

Plotting a pie chart of different category for visualize percentage wise distribution.



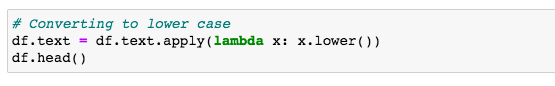


**8. Data Pre-processing**

After data collection, next step is to pre-process the data before model training. In data pre-processing stage we try to clean our data and make it in standard format. As a pre-processing we use different methods. There are like removing double quotes, semicolon, full stop, special characters etc. Also convert all words to lowercase to make it standardize. We have applied below data pre-processing techniques.

**8.1 Converting to lower case:**

All converts all words to lower case to avoid the duplication. And the reason for that is “Cricket” and “cricket” will be considered as 2 separate words if we do not make those two is same format.



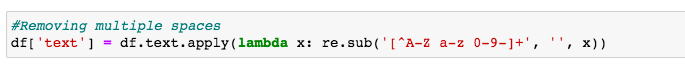
**8.2 Contraction to Expansion:**

Contractions are shortened words, example can’t and don’t. Expanding such words to “can not” and “do not” helps to standardize text. We use the contractions module to expand the contractions.



**8.3 Removing multiple spaces:**

We have used standard python regular expression to remove multiple spaces from document.



**8.4 Removing special character:**

Punctuation is basically the set of symbols [!”#$%&’()\*+,-./:;<=>?@[\]^\_`{|}~]. We remove all punctuations.



**8.5 Removing Stopwords:**

Stopwords are the most common words in English language. Few examples are “the”, “a”, “me”, “is”, “to”, “all”,. These words do not carry much information. Thus we generally removes those words. Here we used Natural Language Toolkit (nltk) to remove stop words from documents.



**8.6 Removing duplicate:**

Removed un-necessary duplicate rows from data using data frame drop duplicate() function.



**8.7 Applying label encoder in news category:**

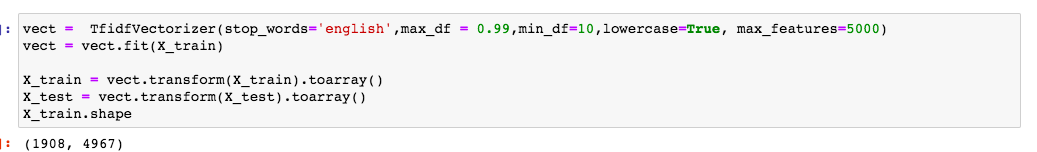
All ML algorithms understand the numbers and not texts. Because of that, all the “text” columns must be converted into “numerical” columns to make it understandable for the algorithm. Lebel Encoding is such a methods for encoding the categorical columns into numerical columns. Here we have used LabelEncoder() module from *Scikit-Learn* library to convert our news category into numeric form.

After applying label encoder here is the news category mapping: {'business':0, 'entertainment':1, 'politics':2, 'sport':3, 'tech':3}



**9. Feature selection**

Here we have used TFIDF techniques for feature selections. Term Frequency-Inverse Document Frequency (TF-IDF) [10] is a techniques to transform text into a numeric format. TFIDF reflects the importance of a word in a document. In text mining we use it as a weighting factor. If the number of occurrence of a particular word in a document increases then the tfidf value will increase. But at the same time it divides that value with the number of documents in the corpus that contain the word. That makes the balance for those words which occurs very frequent. Nowadays TFIDF is mostly used feature selection method.

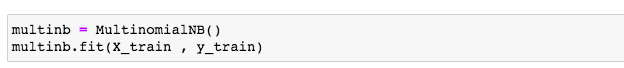


**10. Machine Learning Models**

We have used here two machine learning algorithms to build model.

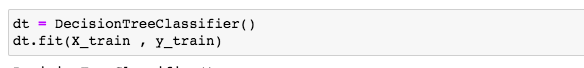
**10.1 Naive Bayes classifier:**

Naive Bayes classifier is a machine learning classification algorithm. It is based on Bayes’ Theorem. In Naïve Bayes we assume that every feature is independent to each other. By applying Naïve Bayes classification techniques we can get reasonable accuracy.



**10.2 Decision Tree:**

Decision Tree is a supervised machine learning techniques. We can use Decision Tree algorithm for both regression and classification problem. Internally it forms a tree like structure where each branches represents weight and each leaf represents different classes. We can train a Decision Tree model using some labelled data.



**11. Model Training:**

Before train each machine learning model we split the entire dataset into train and test data set. And we trained our model using those training datasets.

**12. Evaluation and testing**

After successful training we tested the model using test dataset. We created confusion matrix for each model output and calculated the accuracy score, precision and recall value and also f1 score.

**Confusion Matrix**

To evaluate our classifier performance, we use an important tool called Confusion Matrix. Confusion Matrix gives us a clear picture of the performance of different classifier. It creates a matrix where we can find the number of positive and negative hits of each labels. It gives a comparative analysis in between actual value and predicted value.

Confusion matrix we can use to compute below parameters:

Accuracy

Precision

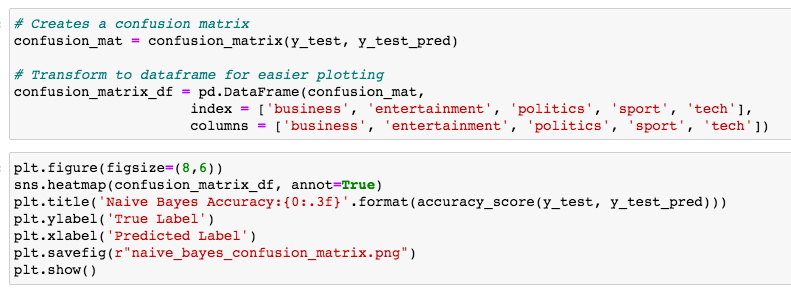
Recall

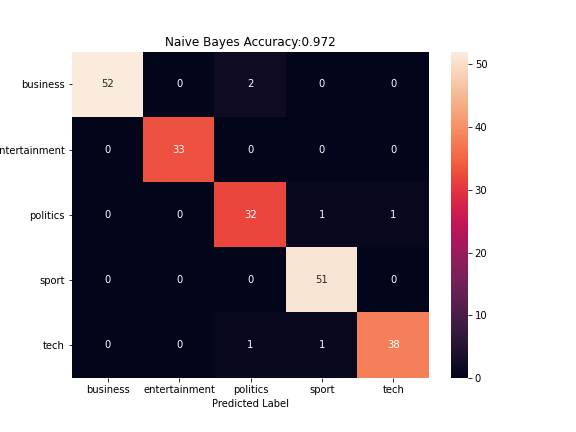
F1-Score

**13. Results:**

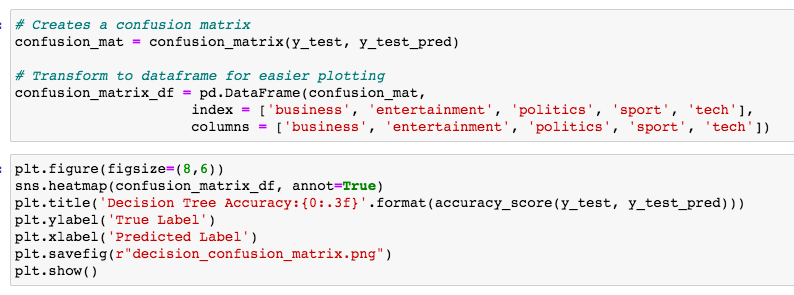
Below is the result for each algorithms (Naïve Bayes and Decision tree).

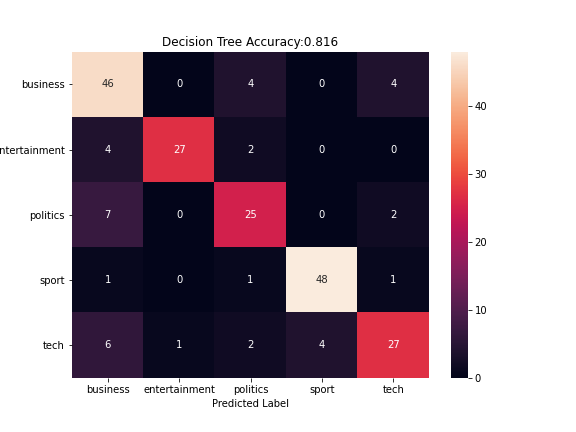
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ML Algorithms | Accuracy | Recall | Precision | F1 score |
| Multinomial Naive Bayes | 97.16 | 97.08 | 97.01 | 97.02 |
| Decision Tree | 81.13 | 80.04 | 81.99 | 80.60 |

****

****

**Multinomial Naive Bayes Confusion Matrix**

****

****

**Decision Tree Confusion Matrix**

**14. Comparison:**

From the above table we can see that the performance of Multinomial Naive Bayes is better than the performance of Decision Tree. So for our final classification module we can use Multinomial Naive Bayes algorithm instead of decision tree. And the accuracy for Multinomial Naive Bayes we are getting here 97.16%.

**15. Conclusion:**

Here we reviewed few well known Machine Learning algorithms. We examined necessary steps which includes data pre-processing, feature extraction, and news category model building in detail. Apart of that we used tfidf based stop words removal approach. In future we can test all those algorithms on larger corpora. Moreover by improving the algorithms we can increase the efficiency of categorization. A combination of algorithm can be also used.

**16. Recommendations and Future Scope:**

In future we can try other Machine Learning algorithms to validate and get more better result.

**17. References:**

[1] Dataset Source: http://mlg.ucd.ie/datasets/bbc.html. Cleaned up version exported to https://storage.googleapis.com/dataset-uploader/bbc/bbc-text.csv

[2] Drury, B., Torgo, L., and Almeida, J.J. ― “Classifying News Stories to Estimate the Direction of a Stock Market Index”. Paper presented at the Information Systems and Technologies (CISTI) 6th Iberian Conference, 2011

[3] M. W. Pope, "Automatic Classification of Online News Headlines," 2007. School of Information and Library Science of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Science in Information Science (November 2007).

[4] L Xin, R Gao, and L Song. “Internet News Headlines Classification Method Based On The N-Gram Language Model”. International Conference on Natural Language Processing and Knowledge Engineering, 2012.

[5] Dr. R. R. Deshmukh, Mr D. K. Kirange. “Classifying News Headlines for Providing User Centered ENewspaper Using SVM”. International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)

[6] D. K. Kirange, R. R. Deshmukh, "Emotion classification of news headlines using SVM" Asian Journal of Computer Science and Information Technology, pp. 104-106, 2012

[7] Dilrukshi, I., De Zoysa, K., Caldera, A, Twitter news classi cation using svm. In: Computer Science & Education (ICCSE), 2013 8th International Conference on, IEEE (2013) 287-291

[8] Seyyed Mohammad Hossein et al “A Novel Text Mining Approach Based on TF-IDF and Support Vector Machine for News Classification” 2nd IEEE International Conference on Engineering and Technology (ICETECH), 17th& 18thMarch 2016.

[9] D. Greene and P. Cunningham. "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering", Proc. ICML 2006

[10] https://en.wikipedia.org/wiki/Exploratory\_data\_analysis

[11] https://scikit-learn.org/stable/