A MINI PROJECT REPORT

On

BBC News Text Classification Using Support Vector Machine

Submitted in partial fulfillment of the requirement of
University of Mumbai for the Course

Natural Language Processing
In
Computer Engineering (VIII SEM)

Submitted By
Aditya Joshi
Ajinkya Darshane

Subject Incharge
Prof. Mayuri Jain

Department of Computer Engineering
A. P. SHAH INSTITUTE OF TECHNOLOGY
THANE – 400 615
UNIVERSITY OF MUMBAI
Academic Year 2019 – 20

Department of Computer Engineering
A. P. Shah Institute of Technology
Thane – 400 615

CERTIFICATE

This is to certify that the requirements for the project report entitled '**BBC News Text Classification Using Support Vector Machine**' have been
successfully completed by the following students:

Name	**Roll No.**
Aditya Joshi	25
Ajinkya Darshane	11

in partial fulfillment of the course Natural Language Processing in Computer Engineering (VIII SEM) of Mumbai University in the Department of Computer Engineering, A. P. Shah Institute of Technology during the Academic Year 2019 – 20.

External Examiner

_____ _____

(Prof. Mayuri Jain)

Subject InCharge

Date:

Place: Thane

Department of Computer Engineering
A. P. Shah Institute of Technology
Thane – 400 615

PROJECT APPROVAL

This project entitled “BBC News Text Classification Using Support Vector Machine” by Aditya Joshi , Ajinkya Darshane is approved for the course Natural Language Processing in Computer Engineering (VIII sem) of Mumbai University in the Department of Computer Engineering.

Subject InCharge:

_____

Date:

Place: Thane

Department of Computer Engineering
A. P. Shah Institute of Technology
Thane - 400 615

DECLARATION

We declare that this written submission for Natural Language Processing mini project entitled "BBC News Text Classification Using Support Vector Machine" represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any ideas / data / fact / source in our submission. We understand that any violation of the above will cause disciplinary action by the institute and also evoke penal action from the sources which have not been properly cited or from whom prior permission has not
been taken when needed.

Project Group Members:

Aditya Joshi

_____

Ajinkya Darshane

_____

Date:

Place:

Table of Contents

Abstract

As in today's world a large number of data is available, classification of data into various sub domains becomes very important. This is required for categorization of data. As data is available in large if is necessary to categories it. A lot of textual data is available and depending upon the context it can be categorized into a particular domain. News data is available in large and it is necessary to categorize news into particular domains. This is needed for user convenience as user can have interest in only a particular section or domain or types of news. Therefore, news classification is necessary. In this report we are focusing on how to use a support vector machine (SVM) for classification of text data into categories. The dataset consists of news that are labeled into 5 categories and we need to train our model to predict the category of our test input news. For this natural language processing is used for processing the sentences before passing them to the support vector machine model for training. We also have a graphical user interface made using tkinter, python where you provide the test sentence and the predicted category is displayed.

Chapter 1

Introduction

1.1 Fundamentals

As our model uses news sentences to predict the category, first we need to perform natural language processing on our text input. We first remove the stop words using package names “stopwords” available in nlp module. As the code is written in Python, for GUI tkinter is been used. The model svm is imported from sklearn library for classification of our news text data. For prediction the words are encoded using Label Encoder in sklearn.

1.2 Objective

Information classification has always been an important topic of research in NLP, as the huge amount of data is available over the internet and classifying this large data is very important and is very difficult task. The Objective of this project is to implement classification of text data. Main task is to preprocess words in sentences and form a dictionary before vectorizing the words. After encoding we pass it to our model to predict classification

1.3 Scope

The scope of the project is to classify the input news into one of the 5 different categories. The model should give a perfect classification of input news. The model will be provided with input news from the entry box present in the graphical user interface and the model predicts the news category. The result is displayed on screen.

Chapter 2

Literature Survey

2.1 Introduction

The IEEE papers have been referred for this project.The first paper referred proposes a method of using Support vector Machine(SVM) algorithm for text classification .In this paper, TF-IDF algorithm was used to classify news articles in Bahasa Indonesia. This algorithm counts the weight of each word with respect to its repetition in the text and the number of files in which it exists. When a word is repeated too many times in all the texts, it means that that word is not important, and that a high precision has been achieved in classification.

The second paper referred to proposes a Text classification approach.It states that text classification is the process of automatically classifying unknown text by suggesting the label that is most likely to belong to it. Some text classification applications are news groupings, document organization, and spam email filtering. Some of the methods that can be used for text classification are k-Nearest Neighbor (KNN), Naive Bayes, Decision Tree, and SVM. We are using this paper to get more information on text classification.

2.2 Literature Review

The first paper proposes a novel method for text classification using the TF-IDF algorithm.The paper suggests that The Bayesian algorithm is a simple and efficient method used in text classification. However, it is not highly efficient because it does not model texts well, nor does it provide a good feature selection. In addition, there are other problems associated with this algorithm. This paper made some modifications to the Bayesian algorithm in order to improve its efficiency. Finally, the algorithm was used to group together the filtered spam messages.A new supervised queue selection method for developing the similarity between a word and a class was introduced to improve the efficiency of text classification.SVM algorithm was used for actual training of model and prediction.

The second paper focuses on supervised learning algorithm SVM.It states that the SVM is a supervised learning which uses the statistical approach. SVM is mainly used to solve regression and classification problems. SVM was originally proposed to solve binary classification The idea of SVM is stated as SVM has the basic idea of implicitly mapping training data into a high-dimensional feature space. A hyperplane is constructed in this feature space that maximizes the separation margin between the hyperplane and the points located closest to it as a supporting vector. Kernel function is a mapping function that mapping the input space into a high dimensional feature space. The most commonly used kernel functions are the Polynomial, Sigmoid, and Radial Basis Function (RBF) .

2.3 Literature Summary

SN	Techniques	Author & Year of Publication	Advantages and Disadvantages
1.	A Novel Text Mining Approach Based on TF-IDF and Support Vector Machine for News Classification	Seyyed Mohammad Hossein Mohammad Shirzad Araghi Morteza Mastery Farahani 2016	Advantages:Efficient algorithm for Text classification using TF IDF. ans SMV
2.	Influence of Word Normalization and Chi-squared Feature Selection on Support Vector Machine (SVM) Text Classification	Ardy Wibowo Haryanto Edy Kholid Mawardi Muljono 2018	Advantages:.Detailed approach of SVM algorithm and implementation.

Table 2.3 Literature survey summary

Chapter 3

Implementation Details

3.1 Overview

The proposed system categories the news according to its type. The application has graphical user interface window which will ask the user to input the text news sentence. This news sentence will be sent for preprocessing and encoding. The output will be then passed to the machine learning model SVG for prediction. The predicted results are returned and displayed on graphical user interface.

3.1.1 Existing Methodology and Systems

Many classification systems exist which use multiple approaches to solve the classification problem. Different algorithms give different accuracy and use different methods for training. The SVG model we propose uses hyperplanes to classify the data points in space. Encoding of data can also be done in multiple ways where different systems use glove vectors or other encoding techniques instead of Tf Idf.

3.1.2 Proposed Methodology and System

We propose to develop a system which can classify the input text news into one of the mentioned categories with maximum accuracy. The user needs to input the test news sentence in the dialogue box of out graphical user interface and the model will provide the accurate category of news.

3.2 Implementation Details

The proposed model is completely written in python programming language. The important domains used are Natural Language Processing (NLP), tkinter in python and machine learning algorithm Support Vector Machines (SVM).

3.2.1 Methodology

The user is asked to input the news sentence which he wants to test in the dialogue box. Our model will get the input sentence and preprocess it before encoding and passing it to the model. The model will predict the result and decoded output is displayed on the screen.

The process of training the model starts by first splitting the data into training dataset and testing dataset. The training dataset consists of 1780 data elements and the testing dataset has 445 data elements randomly created. The next step involves converting the words in lowercase and then using punkt package in nltk for tokenizing the words. The words are then lemmatized using the wordnet package in nltk and entered into a new column so that the data can be easily accessed.

The new column is now onwards used for training and testing of data. As the text data cannot be directly passed to model, first the data is encoded using Label Encoder present in nltk. The encoder encodes the categories from 0 to 4. The Tfidf present in sklearn feature extraction is used to encode the words from our new column. The encoded data is then passed to our model of SVG for training. The trained model gives its accuracy after testing it with a test dataset. When the best accuracy of the model is achieved, we save the model weights so that we don't have to train the model every time we predict results. The inverse transform function of the label encoder is called to decode the predicted results. This decoded result is passed to tkinter to display it to the user.

3.2.2 Details of packages, data set

The dataset used by us for our application is BBC News Dataset. This dataset is freely available on Kaggle website and through google storage api. The dataset consists of 5 categories sports, entertainment, politics, tech and business. The dataset consists of two columns in which the first column gives the category while the second column gives the news corresponding to it. It has a total of 2225 entries.

The packages used are pandas to read the dataset in csv format, tkinter for creating the interface, sklearn for our SVG model of support vector machine. Nltk module in python is used in which packages like word tokenize are used to convert sentences into words and package stopwords are used to form the punctuations and stopwords like the, from, etc. from the list of words. WordNet in nltk is used to create new columns of lemmatized words and Tf Idf is used to vectorize the words. The label encoder module from sklearn package is used for encoding and decoding of our results. The below table lists the data split into categories.

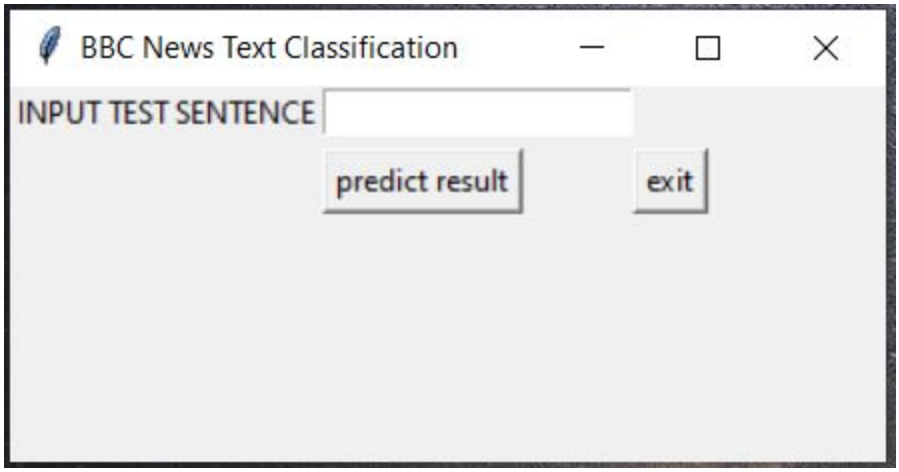
Category	Number of entries
Sports	511
Business	510
politics	417
tech	401
entertainment	386

Chapter 4

Project Inputs and Outputs

4.1 Inputs Details

Below is the basic Interface that user will see:



Here the user will enter the news text sentence that he wants to classify and then the user will press the predict result button to get the results.

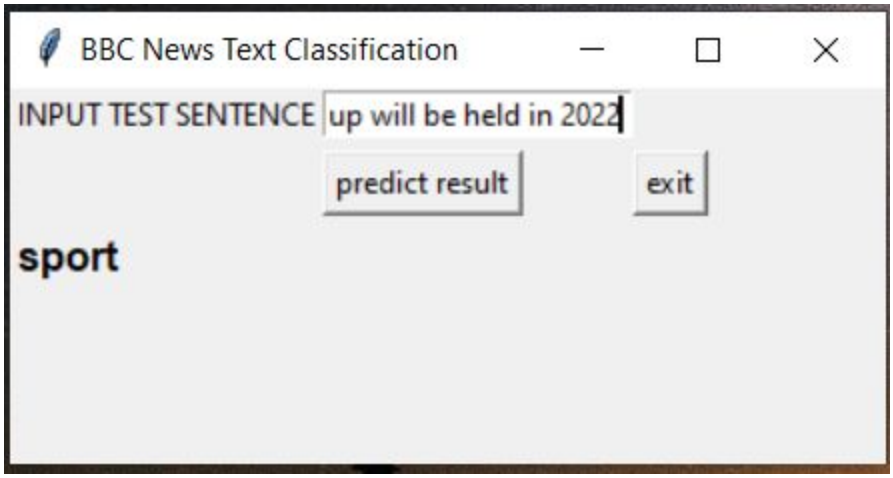
4.2 Evaluation Parameters Details

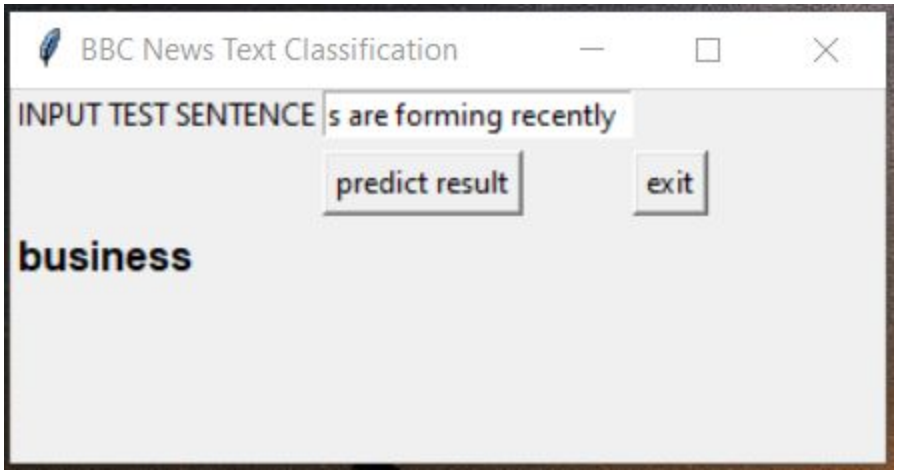
The proposed system can be evaluated by how accurate it is displaying the output of classification.

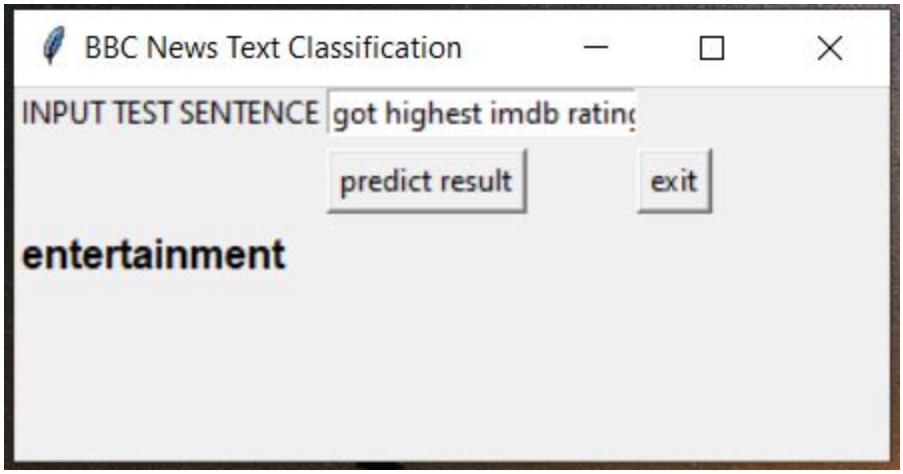
The proposed system is evaluated based on the accuracy of the model and prediction results

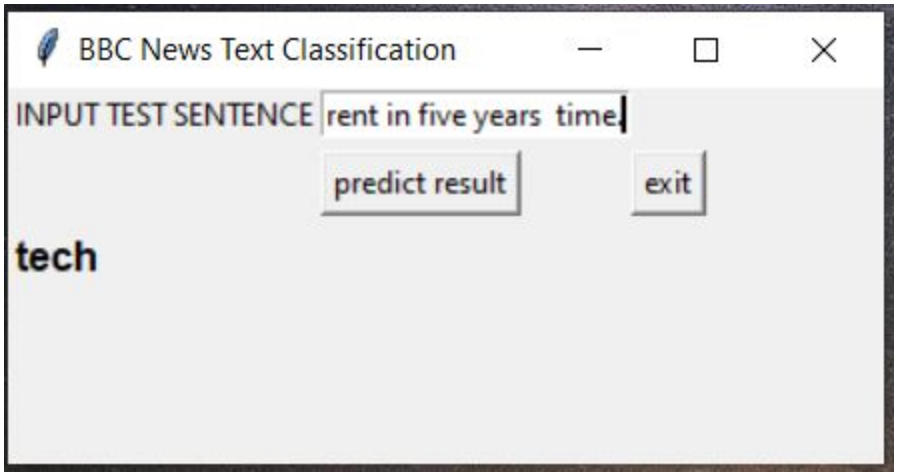
4.3 Output Details and Screenshots

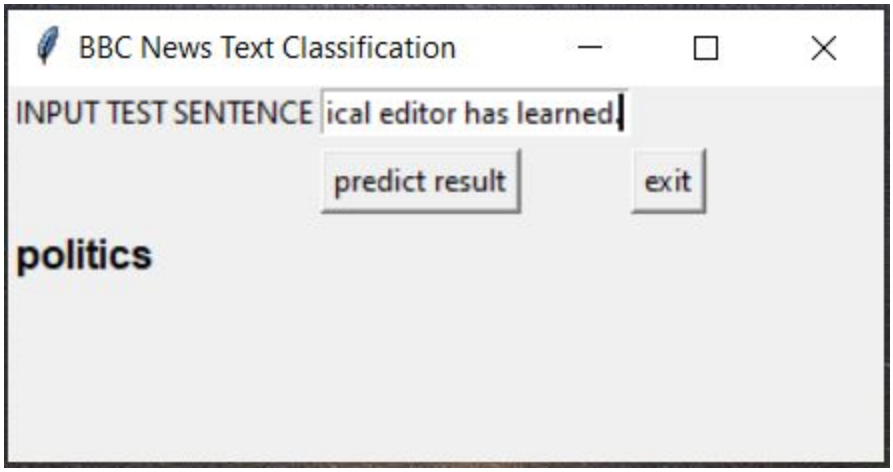
These are some screen shots of news sentences classified into their respective domains.




BBC News Text Classification
INPUT TEST SENTENCE
s are forming recently
predict result
exit
business


BBC News Text Classification
INPUT TEST SENTENCE
got highest imdb ratin
predict result
exit
entertainment


BBC News Text Classification
INPUT TEST SENTENCE
rent in five years time.
predict result
exit
tech


BBC News Text Classification
INPUT TEST SENTENCE
ical editor has learned.
predict result
exit
politics

Chapter 5

Summary and Future Scope

5.1 Summary

The proposed model is very helpful in classification of news into different categories. The model can give classification using 97.64% accuracy. Any sentence of any length can be passed to model as news text input and it will classify it to one of the 5 domains.

5.2 Future Scope

The system has restrictions of only 5 categories of news domain classification. With large dataset with more categories, we can develop a system which can classify any news into multiple domains of news. The model can have more accuracy by training data with different machine learning algorithms and neural networks. Depending on how large data is available, we can also use deep neural networks to train the model to get highest accuracy. We can also develop graphical user interfaces and add more features to our application.

References

[1] M. I. Rana, S. Khalid, and M. U. Akbar, "News classification based on their headlines: A review," in IEEE 17th International Multi-Topic Conference (INMIC), 2014, pp. 211-216.

[2] R. Deshmukh and M. D. Kirange, "Classifying news headlines for providing user centered e-newspaper using SVM," in International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) , 2013, vol 2, Issue 3.

[3] V. Bijalwan, V. Kumar, P. Kumari, J. Pascual, "KNN Based Machine Learning Approach for Text and Document Mining", International Journal of Database Theory and Application, vol. 7, no. 1, pp. 61-70, 2014.

[4] S. B. Mangal and V. Goyal, "Text News Classification System using Naïve Bayes Classifier", International Journal of Engineering Sciences, vol. 3, 2014.

[5] M. W. Pope, "Automatic classification of online news headlines," University of North Carolina at Chapel Hill, november 2007.