A Mini Project Report on

"SVM on News Classification"

Submitted By

Shubham Kokane (BECOMPA-35)
Shreyas Soni (BECOMPA-68)
Piyush Wadi (BECOMPA-72)

(Final Year Computer Engineering)

Guided By

Dr. Archana Chaugule

PIMPRI CHINCHWAD EDUCATION TRUST
A Trusted Brand in Education Since 1990..

Department of Computer Engineering
Pimpri Chinchwad College of Engineering and
Research, Pune 412101 [2020-21]

PIMPRI CHINCHWAD EDUCATION TRUST'S

Pimpri Chinchwad College of Engineering and Research, Pune 412101

CERTIFICATE

This is to certify that *(Shubham Kokane, Shreyas Soni, Piyush Wadi) has* successfully completed the project entitled “SVM on News Classification” in the fulfillment of B. E. (Computer Engineering) LP-III and this work has been carried out in my presence.

Date :

Place :

Dr. Archana Chaugule
Guide and HOD
(Dept. of Comp. Engg.)
Pimpri Chinchwad College of Engineering
and Research, Pune 412101

External

Prof. Dr. Tiwari H.U.
Principal,
Pimpri Chinchwad College of Engineering
and Research, Pune 412101

ACKNOWLEDGEMENT

This is a great pleasure & immense satisfaction to express my deepest sense of gratitude & thanks to everyone who has directly or indirectly helped me in completing my project work successfully. I express my gratitude towards the Project guide (Dr. Archana Chaugule) , Head of Department of Computer Engineering, Pimpri Chinchwad college of Engineering and Research, Pune 412101 who guided & encouraged me in completing the project work in scheduled time. I would like to thank our Principal (Prof. Dr. Tiwari H.U), for allowing us to pursue my project in this institute. No words are sufficient to express my gratitude to our parents for their unwavering encouragement. We also thank all friends for being a constant source of my support.

Shubham Kokane
Shreyas Soni
Piyush Wadi

ABSTRACT

A support vector machine (**SVM**) is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an **SVM** model sets of labeled training data for each category, they're able to categorize new text. We are using SVM to classify News articles dataset and are predicting which category a particular article belongs to. We take the news articles and preprocess them by stemming,lametization and removing the stopwords. After that we apply TF-IDF on it and SVM to classify the news.

INDEX

List Of Figures

1. INTRODUCTION

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called support vectors, and hence the algorithm is termed as Support Vector Machine.

We are using the SVM algorithm to classify and predict the class of the news. The dataset has fields- **Text Feature extraction techniques like TF-IDF.**

TF-IDF:

TF-IDF is a statistical measure that evaluates how relevant a word is to a document in a collection of documents. This is done by multiplying two metrics: how many times a word appears in a document, and the inverse document frequency of the word across a set of documents.

The following figure shows the distribution of data :

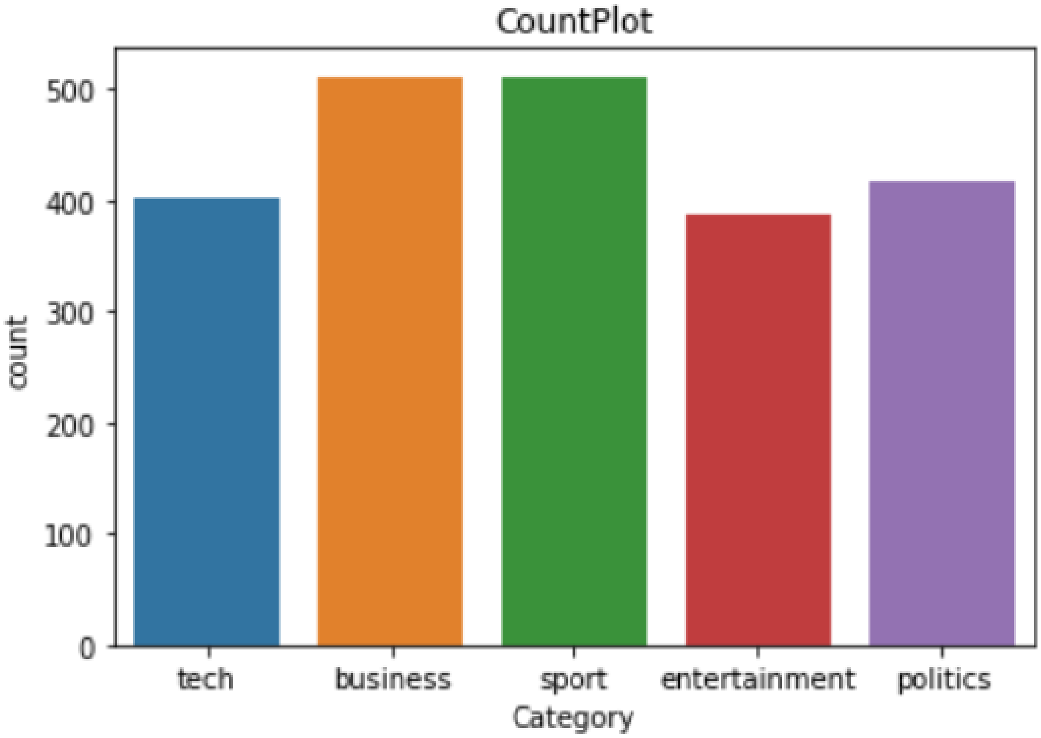


Fig 1.1 Data distribution

2. MOTIVATION

A Support Vector Machine models the situation by creating a *feature space*, which is a finite-dimensional vector space, each dimension of which represents a "feature" of a particular object.

The goal of the SVM is to train a model that assigns new unseen objects into a particular category. It achieves this by creating a linear partition of the feature space into two categories. Based on the features in the new unseen objects (e.g. documents/emails), it places an object "above" or "below" the separation plane, leading to a categorisation (e.g. will purchase or will not purchase). This makes it an example of a non-probabilistic linear classifier. It is non-probabilistic, because the features in the new objects fully determine its location in feature space and there is no stochastic element involved.

Hence, we are using SVM to classify and predict data from our Social Media Ads dataset. We are using SVM to classify News articles dataset and are predicting which category a particular article belongs to. We take the news articles and preprocess them by stemming,lametization and removing the stopwords. After that we apply TF-IDF on it and SVM to classify the news.

3. PROBLEM STATEMENT

Apply the Support vector machine for News classification. Divide data into training and testing data and plot confusion matrices, Classification reports and SVCs for the same. Predict the class of a news article.

4. OBJECTIVES

Objectives of the project are:

- Classify the dataset bbc-news.
- Train and test the data and plot Confusion Matrices.
- Show classification report for the same.
- Take input of a news article and predict the class of the news.

5. SOFTWARE AND HARDWARE REQUIREMENTS

5.1 SOFTWARE REQUIREMENTS:

Python

Pandas

Seaborn

Matplotlib

scikit-learn

Numpy

NLTK

5.2 HARDWARE REQUIREMENTS:

1. Hard Disk: minimum of 1GB of available hard disk space.

2. Memory: recommended 4GB

3. Operating System: *Windows / Ubuntu.*

Tool: Introduction

1.Python:

Python is a widely used general-purpose, high level programming language. It was created by Guido van Rossum in 1991 and further developed by the Python Software Foundation. It was designed with an emphasis on code readability, and its syntax allows programmers to express their concepts in fewer lines of code.

Python is a programming language that lets you work quickly and integrate systems more efficiently.

There are two major Python versions: **Python 2 and Python 3**.

2.Pandas:

What is Pandas?

Pandas is a Python library used for working with data sets.

It has functions for analyzing, cleaning, exploring, and manipulating data.

The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008.

Why Use Pandas?

Pandas allows us to analyze big data and make conclusions based on statistical theories.

Pandas can clean messy data sets, and make them readable and relevant.

Relevant data is very important in data science.

What Can Pandas Do?

Pandas gives you answers about the data. Like:

- Is there a correlation between two or more columns?
- What is average value?
- Max value?
- Min value?

Pandas are also able to delete rows that are not relevant, or contains wrong values, like empty or NULL values. This is called *cleaning* the data.

3.Seaborn:

Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas data structures.

Seaborn helps you explore and understand your data. Its plotting functions operate on data frames and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots. Its dataset-oriented, declarative API lets you focus on what the different elements of your plots mean, rather than on the details of how to draw them.

4.Matplotlib

Matplotlib is one of the most popular Python packages used for data visualization. It is a cross-platform library for making 2D plots from data in arrays. Matplotlib is written in Python and makes use of NumPy, the numerical mathematics extension of Python. It provides an object-oriented API that helps in embedding plots in applications using Python GUI toolkits such as PyQt, WxPython Tkinter. It can be used in Python and IPython shells, Jupyter notebook and web application servers also.

Matplotlib has a procedural interface named Pylab, which is designed to resemble MATLAB, a proprietary programming language developed by MathWorks. Matplotlib along with NumPy can be considered as the open source equivalent of MATLAB.

5. Scikit-learn

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistent interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.

6. Numpy

NumPy is a Python package. It stands for 'Numerical Python'. It is a library consisting of multidimensional array objects and a collection of routines for processing of arrays.

Numeric, the ancestor of NumPy, was developed by Jim Hugunin. Another package Numarray was also developed, having some additional functionalities. In 2005, Travis Oliphant created the NumPy package by incorporating the features of Numarray into the Numeric package. There are many contributors to this open source project.

7. STEPS OF IMPLEMENTATION WITH OUTPUT

7.1 Steps of implementation :

1. Load Dataset of BBC-news.
2. Preprocess the data
3. Split data into training and testing data
4. Scale the data
5. Load data into SVM and train the data
6. Plot Confusion matrices
7. Show classification reports

7.2 Output Screenshots

1. Class Prediction Error

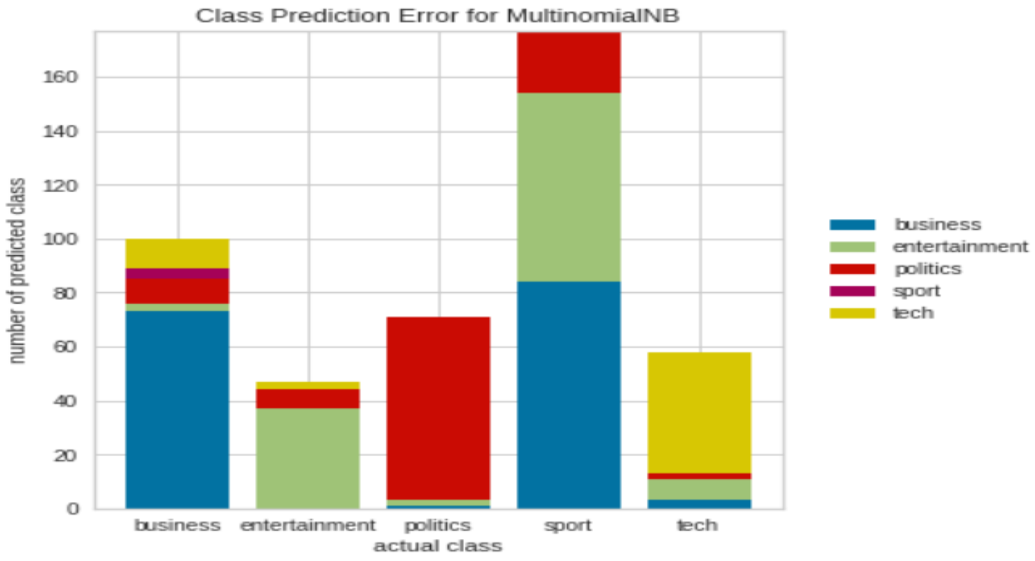
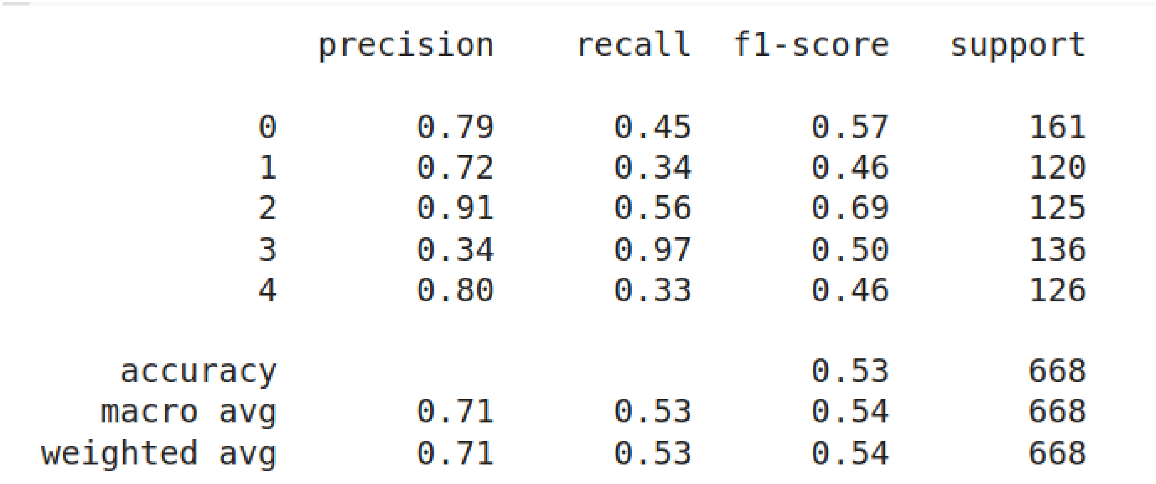


Fig 7.2.1 Class Prediction Error for MultinomialNB

2. **Classification Report**



```
              precision    recall  f1-score   support

           0       0.79      0.45      0.57       161
           1       0.72      0.34      0.46       120
           2       0.91      0.56      0.69       125
           3       0.34      0.97      0.50       136
           4       0.80      0.33      0.46       126

    accuracy                           0.53       668
   macro avg       0.71      0.53      0.54       668
weighted avg       0.71      0.53      0.54       668
```

Fig 7.2.2 Classification Report

8. CONCLUSION

Hence, using SVM we have classified news articles dataset and we have also got classification results. We can see the F1 score of the SVM model in the results which gives us an accuracy of 53 percent.