North South University

Department of Electrical & Computer Engineering

PROJECT REPORT

Course Code: CSE445

Course Name: Machine Learning

Section: 03

Project Name: News classification based on similarity

Submitted To: ITN

Submission Date: 25.05.21

Student Name	Student Id
Ankur saha	1620753642
Md. Sadman Sakib	1620676042
Majharul Alam Evan	1621460642

Abstract

Social site, Web blogs, many news are growing exponentially in the big data area. Even though there is a lot of news to explore, there are challenges to finding news That meets the users' interest. Now-a-days, News providers used to share their news headlines in various web sites and web blogs. Also, there are many news groups whom share their news headlines in micro blogging services. These data may carry out much valuable information which will relevant to many social research areas. Thus, the purpose of this research is to classify news. This project proposes a news content classifier that classifies text news content that sets predefined labels. As we know, the short messages were extracted from micro blog. Each short message was classified manually. These classified data were used to train the machine learning techniques. The data were trained using SVM (Support Vector Machine) machine learning techniques. The main reason of using SVM for the current study is, SVM supports high dimensional data. Current research is a high dimensional problem as a large number of features will be collected using short messages. Cross validation was done in order to avoid the biasness of data. The performance of the system will be the effectiveness of the system. Thus precision and recall values are calculated to measure the performance of the system. The results show that the system provides high performance. Keywords-category predictor; SVM; accuracy

Introduction

THE news world is bustling every second with news entering from all sources. There are multiple news channels, online news portals that let out the daily proceeding every minute. Different types of news make it to these portals. Whether it is print media or electronic media, news stories are important and flowing everywhere. So, it is important to have an efficient system of segregating news into different categories. Technology can be used to enhance and better this system by the proper use of machine learning . The news headlines will be used to train the model and later the machine will be able to predict the category of the news item very fast and accurate. This will be helpful for all news channels and apps as it will give them an efficient and speedy way to do their job. Large data can be segregated easily which is a very good thing for them. We have a dataset from Kaggle that have seven columns, one has the news category and the others contain headline , authors, link , short description , date. There are more than 200k rows in the data set. In this work, one supervised machine learning algorithms, Support Vector Machine with the similar categories are predicted .

The news category predictor aims to recognize and categorize different articles based on content/information type. The automatic news classification plays a vital role in processing a massive amount of articles. It can classify and label the news articles by analyzing the content (i.e. extracting feature values) to quickly access where they are focused in, allowing efficient and speedy news dissemination. Additionally, news websites can also increase their visibility by developing a recommendation system that suggests/ recommends relevant news to attract more attention. Several performance evaluation of news category predictors using machine learning (ML) algorithms over different datasets.

Methodology

The ultimate aim of this study is to classify the news into specific categories and analyze the performance of the category predictor. Initially, data are collected and preprocessed, then the content of text document (Dj) is converted into useful features (w1j... wkj) by feature extraction algorithms such as unigrams. The extracted features are transformed into numeric data that act as inputs for machine learning algorithms or classifiers (NB and RF). Finally, the ML models are trained on these transformed features, and the performance is evaluated on the test dataset. The research methodology/work flowchart is given in Figure .

A) Dataset

In this study, a BBC-originated news data set was used, which is obtained from Kaggle. It from BBC news website corresponding to stories in several typical areas such as , Politics , wellness , entertainment , sports , arts and culture etc. We focus in total 26 typical areas.The dataset are not balanced . Most of the news are belong to the political areas. The distribution of classes plays an important role in classification, and balanced datasets result in better learning models.

Fig01- class workflow

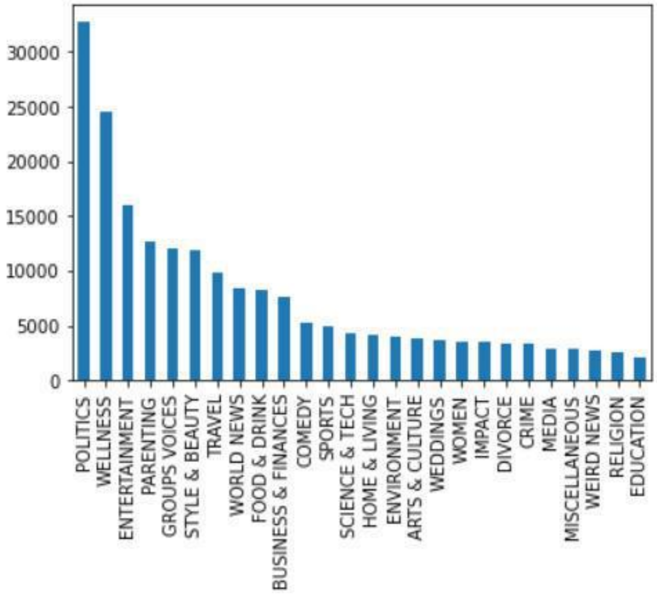


Figure 2 : Class Distribution Bar Chart

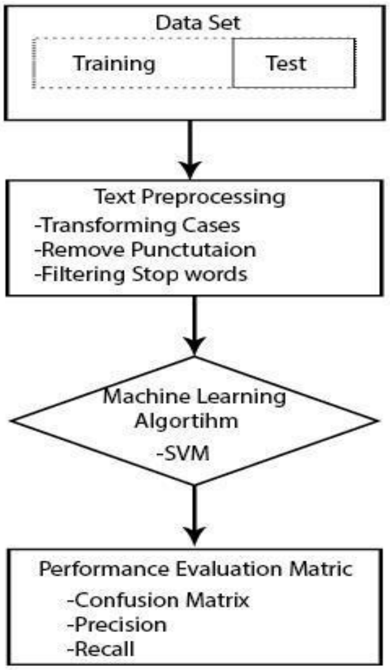


Figure 1: Class Flowchart

B)Text cleaning and Preprocessing:

Text preprocessing, also known as text cleaning, is a preliminary and critical stage in news categorization that information. Stop words, punctuation, special characters, unrelated decreases the amount of space required and improves classification efficiency [18]. The dataset is usually unstructured, with a mix of important and meaningless phrases, quotes, and dates are all unnecessary information that adds no predictive value to the classifier/model. They merely take up space and can cause the ML model to be distorted. To reduce the distortions contributed to the model, a cleaning step should be done before extracting any feature from the raw dataset. Several studies are presented in this publication.

1) **Transforming Test:**
   Text in the same letter size (i.e. lower case) is transformed to exclude homologous words that differ only in case. For example, in a true sense, the words "fruit" and "fruit" are interchangeable and should not be used for prediction.

2) **Removing adj,verb and adverb:**
   Parts of speech such as adjective, verb and adverb are disposed of this process and simplifies computations in the next steps.

3) **Filtering stop words:**
   This strategy is mostly used to eliminate unneeded words or words with no special meaning, such as "the," "an," "a," "what," and similar terms, so that the classifier cannot associate stop words with significant class attributes. Furthermore, neither the most often used nor the least frequently used terms contribute to the predictive power model. As a result, they must be taken out of the training set. We used the nltk package to obtain a list of English stop words, which we subsequently eliminated from the dataset.

Results:

Three assessment metrics were used to evaluate the experimental results: Accuracy, Precision, and Recall. Accuracy: Accuracy tells us how confident the model is in detecting certain things. There are two types of people in our world: those who are positive and those who are negative Precision: We emphasize that precision is important. about the likelihood of generating a right positive the categorizing of classes. Recall is a term that describes how sensitive something is the goal of the model is to find the positive class.

The performance results of multi-class category predictors based on different supervised learning models are evaluated and compared in this section. This study's learning model is SVM. The evaluation was done by observing each category predictor's prediction results by analyzing the Confusion Matrix and quantifying Precision, Recall, and overall Accuracy. This analysis was made on a test dataset consisting of more than 200K news samples.

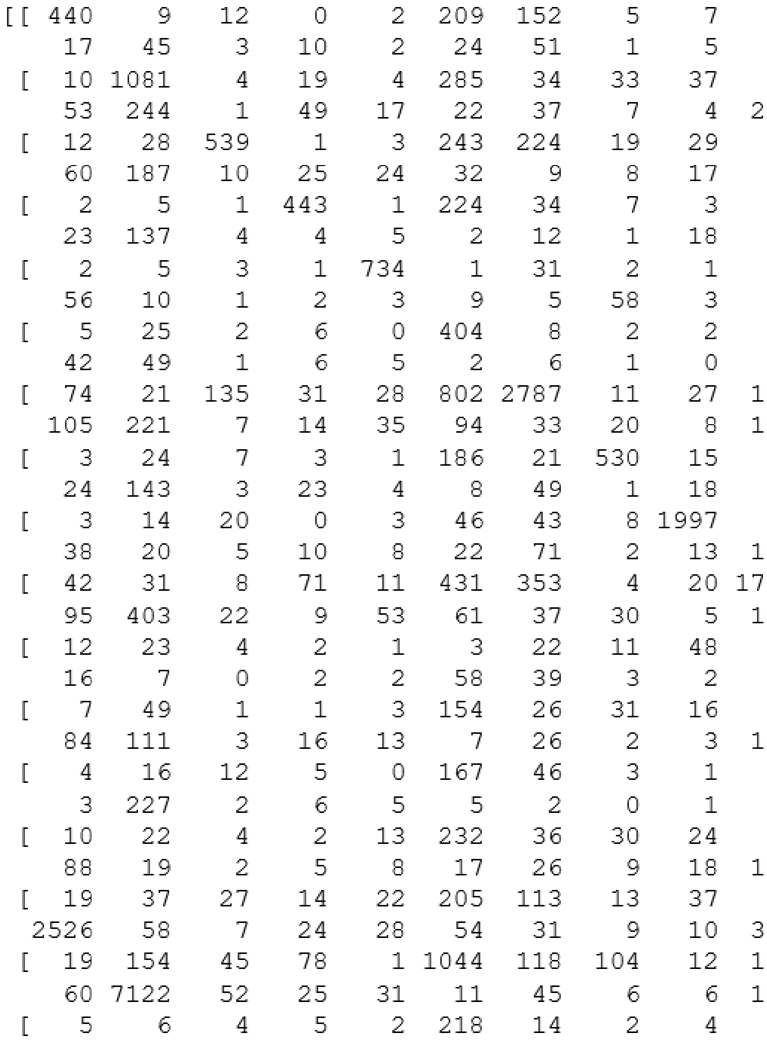
	Precision	Recall	F1-score	Support	Accuracy
Politics	0.55	0.38	0.45	1148	0.60
Parenting	0.57	0.47	0.51	2309	
wellness	0.57	0.33	0.42	1611	
Travel	0.79	0.71	0.75	1033	

Figure 03- Result Table

If we analyze the category predictor model based on SVM algorithm, the prediction performance needs to be more satisfactory. As we get Training accuracy score: 0.60
And Testing accuracy score: 0.77

Confusion Matrix:

The Confusion Matrix obtained from Support Vector Machine are described as:



```
[[ 440    9   12    0    2  209  152    5    7
    17   45    3   10    2   24   51    1    5
 [  10 1081    4   19    4  285   34   33   37
    53  244    1   49   17   22   37    7    4   2
 [  12   28  539    1    3  243  224   19   29
    60  187   10   25   24   32    9    8   17
 [   2    5    1  443    1  224   34    7    3
    23  137    4    4    5    2   12    1   18
 [   2    5    3    1  734    1   31    2    1
    56   10    1    2    3    9    5   58    3
 [   5   25    2    6    0  404    8    2    2
    42   49    1    6    5    2    6    1    0
 [  74   21  135   31   28  802 2787   11   27   1
   105  221    7   14   35   94   33   20    8   1
 [   3   24    7    3    1  186   21  530   15
    24  143    3   23    4    8   49    1   18
 [   3   14   20    0    3   46   43    8 1997
    38   20    5   10    8   22   71    2   13   1
 [  42   31    8   71   11  431  353    4   20  17
    95  403   22    9   53   61   37   30    5   1
 [  12   23    4    2    1    3   22   11   48
    16    7    0    2    2   58   39    3    2
 [   7   49    1    1    3  154   26   31   16
    84  111    3   16   13    7   26    2    3   1
 [   4   16   12    5    0  167   46    3    1
     3  227    2    6    5    5    2    0    1
 [  10   22    4    2   13  232   36   30   24
    88   19    2    5    8   17   26    9   18   1
 [  19   37   27   14   22  205  113   13   37
  2526   58    7   24   28   54   31    9   10   3
 [  19  154   45   78    1 1044  118  104   12   1
    60 7122   52   25   31   11   45    6    6   1
 [   5    6    4    5    2  218   14    2    4
```

Fig04-Confusion matrix

Conclusion and Future works:

A comparison of the prediction performance of the multi-class category predictor is presented in this work. Well-known machine learning methods SVM was used to construct news category predictors. We next evaluated the Confusion Matrix and assessed the test dataset's Precision, Recall, and overall Accuracy using performance assessment criteria. We used the SVM algorithm to get correctly categorizing news to get the accuracy.

REFERENCE:

Link1-
https://www.researchgate.net/publication/345579704_Design_and_Analysis_of_News_Category_Predictor

Link2-

http://www.ijstr.org/final-print/jan2020/A-Comparative-Analysis-Of-News-Categorization-Using-Machine-Learning-Approaches.pdf?fbclid=IwAR3vRxVgIfUdZLbfzmuH7ujxGJoUeZBSTBzu1jq4_kx1OS0VAhwGuUWVu8Q