Nitte Meenakshi Institute of Technology,

Department of Computer Science and Engineering

18CSE751 Introduction to Machine Learning

Learning Activity Proposal

Topic based News classification

Harika H P, Chakrala Navyasree

1. Abstract

News is available in various places like website, television, papers etc. in online and offline modes, which includes politics, business, entertainment, sports and other general categories. Automatic classification of the news articles is very important for fast and effective communication. Text processing and text mining are basic techniques involved in news classification. News classification helps in various emerging applications. This project is focusing on the topic-wise news classification using Artificial Intelligence techniques. Content Mining is the application of mechanized approaches for understanding the information available in the content archives. Text Mining is delineated in a way to assist the business find out essential knowledge from text-based content. With the help of lexical approach and machine learning techniques we can perform news classification to demarcate them into distinct topics for easy deliberation.

A few of the papers from the literature in the context are studied and analyzed for understanding the problem domain and the solution approaches. In our project AI based methods are implemented for doing the classification of news with topic discovery approach. This report presents results of experimenting with a few algorithms for developing models for news classification, viz., simple baseline model, decision tree, Random Forest, multi-nominal Naïve Bayesian, Multi-layered perceptron, and Support vector. Their performances are compared based on precision, recall, F1 score measure. Out of them Bayesian method and MLP are found to be the best ones. Further experiments can be done with deep neural networks.

2. Introduction

❖ Background

The motivation for exploiting background knowledge in text classification is attributed to two reasons. First, more information from texts can make more reasonable classification. Second, people have basic concepts and general knowledge in their mind; however, the common corpora/datasets are some kinds of special case which would lack some basic concepts and general knowledge. These basic concepts and general knowledge are the background knowledge in our life.

- **Brief history of Technology/concept**

There are different steps involved in news classification. Classification is a difficult activity as it requires pre-processing steps to convert the textual data into structured form from the un-structured form. Text classification process involves following main steps for classification of news article. These steps are data collection, pre-processing, feature selection, classification techniques application, and evaluating performance measures.

- **News Collection**

The first step of news classification is accumulating news from various sources. This data may be available from various sources like newspapers, press, magazines, radio, television and World Wide Web and many more.

- **News Pre-processing**

After the collection of news text pre-processing is done. As this data comes from variety of data gathering sources and its cleaning is required so that it could be free from all corrupt and futile data. Data now needs to be discriminated from unrelated words like semicolon, commas, double quotes, full stop, and brackets, special characters etc.

- **Feature Selection**

When there exist a large number of features and each of the features is a well-known descriptive word for each class, a lot of time may be required in classification and it may be possible that expected accuracy may not be achieved and to overcome these issues, a process named as feature selection is adopted in which only those relevant and highly effective features are chosen, which may prove more noticeable for better news classification. A large number of techniques exist in literature for selecting appropriate features like Boolean weighting, Class Frequency Thresholding, Term Frequency Inverse Class Frequency, TF-IDF, Information Gain.

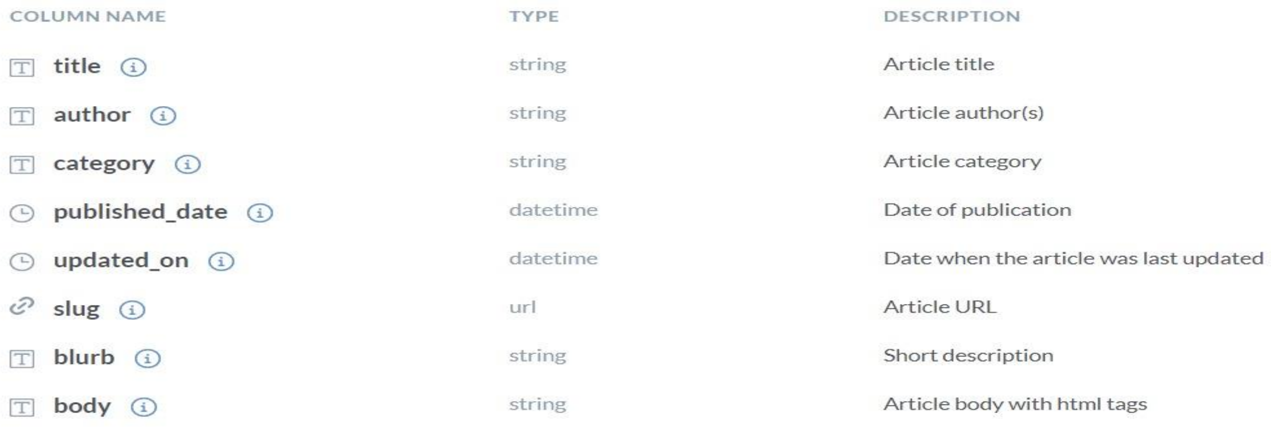
- **News Classification**

After feature selection the next phase is the classification phase which is an important phase in which the aim is to classify the unseen news to their respective categories.

3. Data Set

https://data.world/elenadata/vox-articles

1Vox Media kindly provided this dataset as part of the KDD 2017 Workshop on Data Science + Journalism. It contains all Vox articles published before March 21, 2017.



COLUMN NAME	TYPE	DESCRIPTION
title	string	Article title
author	string	Article author(s)
category	string	Article category
published_date	datetime	Date of publication
updated_on	datetime	Date when the article was last updated
slug	url	Article URL
blurb	string	Short description
body	string	Article body with html tags

4. Machine Learning Methods

- **Naive Bayes:**

Naive Bayes is a probabilistic classifier based on text features. It calculates class labels and probability of classes. Naive Bayes is not made up of a single algorithm for classification but it includes a large number of algorithms that work on a single principal for training classifiers and the principal states that the value of a particular feature is autonomous of value of any other feature specified in a class. In the past classification of news article Naive Bayes were used. But due to its incorrect parameter assessment revamped accuracy was reported. The best thing about Naive Bayes algorithm is that it works equally well on both textual as well as numeric data and it is easy to implement and calculate. But it shows poor performance when the features are correlated like short text classification.

- **Support Vector Machines (SVM):**

SVM has been used a lot for news text classification. SVM has a unique feature that it includes both negative and positive training sets which is generally not preferred by other algorithms.

- **Artificial Neural Networks:**

This network has got the concepts from neurons in which huge calculations are performed very easily by providing sufficient input and are used to estimate functions which are based on large number of inputs. Neural network when used with Naive Bayes presented a new approach known as Knowledge based neural network which is efficient in managing noisy data as well as outliers. Artificial neural network yields good results on complex domains and allows performing fast testing. But the training process is very slow.

- **Decision Trees:**

Decision Trees are quite easily perceived and rules can be easily produced through them. Decision Trees can be used to solve intricate problems very easily. It comes with a clause that training decision tree is an expensive task. Besides this, news can be connected to one branch only. If the mistake occurs at the higher upper level it cause the whole sub tree go invalid.

- **K-nearest neighbors:**

K-nearest neighbors is a simple algorithm and a non-parameterized way of classification and regression in case of pattern recognition. For using this algorithm, we need to refer to K text documents. It reckons the similarity against all documents that exists in the training set and uses it for making decisions about presence of class in the desired category. Neighbor that has same class are the most probable ones for that class and the neighbors with highest probability are assigned to the specific class. K-nearest neighbors is effective and non-parameterized algorithm. The biggest pitfall is that it requires a lot of classification time and it is also difficult to find an optimal value of K.

K-nearest neighbor is a type of lazy learning where function Generalization beyond the data is delayed until a query is made to the system. K-nearest neighbor is one of the simplest machine learning algorithms.

5. Assessment

- **Accuracy**

Accuracy is often the most used metric representing the percentage of correctly predicted observations, either true or false. To calculate the accuracy of a model performance, the following equation can be used:

In most cases, high accuracy value represents a good model, but considering the fact that we are training a classification model in our case, an article that was predicted as true while it was actually false (false positive) can have negative consequences; similarly, if an article was predicted as false while it contained factual data, this can create trust issues. Therefore, we have used three other metrics that take into account the incorrectly classified observation, i.e., precision, recall, and F1-score.

$$\text{Accuracy} = TP+TN/(TP+TN+FP+FN)$$

- **Recall**

Recall represents the total number of positive classifications out of true class. In our case, it represents the number of articles predicted as true out of the total number of true articles.

$$\text{Recall} = TP/(TP+FN)$$

- **Precision**

Conversely, precision score represents the ratio of true positives to all events predicted as true. In our case, precision shows the number of articles that are marked as true out of all the positively predicted (true) articles:

$$\text{Precision} = TP/(TP + FP)$$

- **F1-Score**

F1-score represents the trade-off between precision and recall. It calculates the harmonic mean between each of the two. Thus, it takes both the false positive and the false negative observations into account. F1-score can be calculated using the following formula:

$$F1 = 2 \text{ x } (\text{precision x recall}) / (\text{precision} + \text{recall})$$

6. Presentation and Visualization

The dataset is divided into the training model, development model, testing model.

We classify the given dataset into the different categories like Business and Finance, Health care, Science and Health, Politics and policies, and criminal justices. We used word cloud to visualize the data. For each category of data, we calculate precision, recall, f1-score and accuracy.

7. Roles

Harika: collection of data set and pre-processing of data

Navyasree: developing the model using machine learning algorithms

8. Schedule

Date	**Tasks to be Completed**
12/12/21	selection of the project and the data set
21/12/21	submission of learning activity proposal
05/01/22	50% of project implementation
17/01/22	final project report submission and final presentation

9. Bibliography

[1] Zhenzhong Li, Wenqian Shang, Menghan Yan, “News text classification model based on topic model”, IEEE International Conference on Data Mining Workshops (ICDMW), Volume: 1, Pages: 1375-1380, 2018

[2] Vrusha U.Suryawanshi, Pallavi Bogawar, Pallavi Patil, Priya Meshram, Komal Yadav, Nikhil S. Sakhare, “Automatic Text Classification System”, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 4 Issue 2, February 2015.

[3] Ari Aulia Hakim, Alva Erwin, Kho I Eng, Maulahikmah Galinium, Wahyu Muliady, “Automated Document Classification for News Article in Bahasa Indonesia based on Term Frequency Inverse Document Frequency (TF-IDF) Approach”, 6th International Conference on Information Technology and Electrical Engineering (ICITEE), 2014.

[4] S. M. H. Dadgar, M. S. Araghi and M. M. Farahani, "A novel text mining approach based on TF-IDF and Support Vector Machine for news classification", IEEE International Conference on Engineering and Technology (ICETECH), Coimbatore, India, 2016.

[5] Akshita Bhandari, Ashutosh Gupta, Debasis Das, “Improvised apriori algorithm using frequent pattern tree for real time applications in data mining”, International Conference on Information and Communication Technologies (ICICT), 2014.

[6] Nabamita Deb, Vishesh Jha, Alok K Panjiyar, Roshan Kr Gupta, “A Comparative Analysis Of News Categorization Using Machine Learning Approaches”, International Journal Of Scientific & Technology Research, Volume 9, Issue 01, January 2020.

[7] H. Duong and V. Truong Hoang, "A Survey on the Multiple Classifier for New Benchmark Dataset of Vietnamese News Classification," International Conference on Knowledge and Smart Technology, 2019.

[8] W. Wang, X. Cui and A. Wang, "News Analysis Based on Meta-synthesis Approach", IEEE International Computer Software and Applications Conference, 2008.

[9] U. Suleymanov, S. Rustamov, M. Zulfugarov, O. Orujov, N. Musayev and A. Alizade, "Empirical Study of Online News Classification Using Machine Learning Approaches”, IEEE International Conference on Application of Information and Communication Technologies (AICT), 2018.

[10] Gurmeet Kaur, and Karan Bajaj, “News Classification and Its Techniques: A Review”, IOSR Journal of Computer Engineering (IOSR-JCE), Vol.18, issue 1, 2016