**Introduction:**

This project, News Categorization, is a machine learning model that categorizes news articles based on the news headlines and short summary. It has been observed that such kind of tool or model does not exist – at least not in its entirety. This news categorization model would be beneficial for news editors, owners to appropriately classify the news articles without human intervention.

For long time, this process of categorization news was done **manually** by people and used to allot news to respective section(category). Allotting news to respective section manually is challenging and inelegant tasks, it requires the knowledge about each category and more time for review before finalizing the categories.

With digitalization of newspaper, the news gets updated every moment and allocating to them to appropriate category can be cumbersome task and need more human SMEs in respective section to categories news appropriately and accurately. Manually allotting news category can lead to inaccurate classification that would confuse users or may not find right news in the right category. This will cause user frustration and bad impact on newspaper brand/popularity.

The model developed as part this project using supervised machine learning techniques, that would learn from existing news headlines and short description and predict the news category appropriately.

With the help of this model the news categorization can be automated, and it would save manual efforts and help users to read the news of their interest in right category.

This project has major four stages –

1. Collect the news headlines, short summary and categories for past news.

2. Clean and preprocess these headlines and summary into machine understandable format.

3. Identifying the different ML algorithms suitable for this task and train the ML Models.

4. Evaluating each ML model for its accuracy and finalizing the model with highest accuracy.

**Background:**

From the beginning, since first newspaper printed, every news that makes into page has had a specific section allotted to it. The newspaper style, news sections, format etc. have been changed over the time but not the categorization of the news and it still carried over even into the digital version of newspaper. News articles are not limited to few topics, it covers a wide range of interest from politics to sports to movies and so on.

With transformation of newspaper from printed version to digital version, the categorization was done manually in most of the organization by reading the news headlines and short summary. This kind of task is time consuming and need more human power to classify huge amount of news that gets introduced every moment. With this fast space life, users do not spend time on searching news, they visit the section of their interest and read news. The categorization of the news would help them read the news of their interest by vising respective section, to match the speed of new news generation, human categorization model would not benefit organization grow fast and keep up with the competitors.

To overcome such problem of manually categorizing news by building a machine learning model that would learn from the news headlines and short summary and classify them into appropriate category, this would save human efforts as well as time and speed of news on end user’s favorite section.

Text categorization or classification is a way of assigning documents to one or more predefined categories. This helps news editors, end users to allot and find the appropriate news category.

**The main problem** we are going to handle as part of this project is, **to categorize news based on headlines and short description by using supervised machine learning classifiers**.

**Data Description:**

This dataset contains around 200k news headlines from the year 2012 to 2018 obtained from HuffPost. These news headlines and short descriptions were posted in different newspapers in the past, corresponding news categories were also collected for each news which was published under that category.

Each news headline has a corresponding category. Categories and corresponding article have different counts. These categories are predefined the time of collecting the data and assigned them accordingly.

**Evaluation Of Case:**

To evaluate this problem statement, we have used different methods to explore and preprocess the news data. The exploratory analysis and model evaluation done on the news data explained as below -

**Graphical Analysis:**

**Shape and Sample of Data**: This data set contains 200K records and have category for each headline and short description along with author of news and source link and date it was published.

**A screenshot of a computer

Description automatically generated with medium confidence**

**Number of news per Category:** The below bar chart shows the number of news published under each category. Based on this bar chart it shows that we are dealing with imbalanced classes, which means we do not have same number of news across the news categories and we need to handle this using appropriate techniques such as SMOTE before training the ML model.

A picture containing text, measuring stick, screenshot

Description automatically generated

**Number of News per year**: This bar chart shows number of news were published in each year. We can see the data was collected from news published year 2012 to 2018. Since it was collected in 2018, we do see number of news are less than 10K in that year.

Chart, bar chart

Description automatically generated

**Number of words in headlines**: This histogram shows that distribution of number of words in each headline. It seems there are headlines with more than 25 words and their count is low, these are outliers in data set, and we are going to remove them.

Chart, histogram

Description automatically generated

**Number of words in summary:** Based on this histogram we can see that the distribution is skewed, and outliers exist. We need to remove the outliers with word count more than 50.

Text

Description automatically generated with low confidence

**Dimensionality & Feature Reduction and Feature Engineering:**

The main goal of this project is to predict the appropriate category from news headline and news short description, we are going to club news headlines and short description features into one feature called news text, so these are the two features (news text and category) we need to train model on. Hence, we are going to drop rest unwanted features from the data set.

The below list of features will be dropped from the data set -

1. authors - This feature tells who wrote the news, but it would help use determine the news category, we are going to drop this feature.

2. link - this feature shows source of the news, and it does not require for news categorization. So, it will be dropped.

3. date - this feature explains what date the news was published, and this would not help to solve the problem, hence it will be removed from the data set.

4. year - This feature was created from date, to see the distribution of news and headlines and this will be removed as well.

5. num\_words\_head and num\_words\_summary - These features were created to see the words distribution in headlines and summary respectively and are used to remove outliers from the dataset. Once outliers are removed based on these feature numbers, it will be dropped from data set.

**Model Selection:**

When working on data science project to find insights from the data or appropriately categorize the data based on given inputs, we often need to use machine learning algorithms that helps resolve the problem. There are several machine learning algorithms that can be used but to select appropriate algorithms based on the data and problem, it is crucial step in the project and required information about the algorithms and how it works.

Why model selection is important? In machine learning projects model selection is a process data scientist use to compare the relative value of different machine learning models and determine which one is the best fit for the observed data.

Since, the problem statement of this project falls under classification, we are going to use several different supervised machine learning classifiers to train on news data set and evaluate each one of them for accuracy and several different evaluation metrics.

The first machine learning algorithm selected for classifying categories for news data set is **Naive Bayes Classifier**.

**1. Naïve Bayes Classifier** - It is a classification technique based on Bayes’ Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. It works on Bayes theorem of probability to predict the class of unknown data sets.

**Reason to use Naïve Bayes Classifier** -

- It is easy and fast to predict the class of the test data set. It also performs well in multi-class prediction. When assumption of independence holds, a Naive Bayes classifier performs better compared to other models like logistic regression.

- Because of the class independence assumption, naive Bayes classifiers can quickly learn to use high dimensional features with limited training data compared to more sophisticated methods.

- It is the most applied commonly to text classification. Though it is a simple algorithm, it performs well in many text classification problems.

- It performs well in case of categorical input variables compared to numerical variable(s). Because for numerical variable, normal distribution is a strong assumption.

Based on the advantages of Naïve Bayes Classifier, it helps to resolve our problem statement and since our training data is text (converted to vectors with high dimensions), this algorithm best suits for building model.

**2. Linear Support Vector Machine** – SVMs are a set of supervised learning methods used for classification, regression, and outliers’ detection.

Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier. For multiclass classification, the same principle is utilized after breaking down the multiclassification problem into multiple binary classification problems.

**The advantages of support vector machines are:**

- Effective in high dimensional spaces.

- Still effective in cases where number of dimensions is greater than the number of samples.

- Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.

**3. Logistic Regression** - Logistic regression is a classification algorithm, used when the value of the target variable is categorical in nature. Logistic regression is most used when the data in question has binary output, so when it belongs to one class or another, or is either a 0 or 1. This algorithm can be used for multiclass classification because multiclass classification is built on the binary classification. The approach used for multi-class classification is one vs all method.

**4. XGBoost Classifier** - It is an optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way.

It is more apt for multi-class classification task. By default, XGBClassifier or many Classifiers uses objective as binary but what it does internally is classifying (one vs rest).

Advantages of XGBoost Classifier –

- It is Highly Flexible and supports regularization

- It uses the power of parallel processing and faster than Gradient Boosting

- It is designed to handle missing data with its in-build features.

- It Works well in small to medium dataset

**Model Evaluation:** Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future.

Model evaluation aims to estimate the generalization accuracy of a model on future (unseen/out-of-sample) data.

To evaluate the performance of the model chosen for classifying the news categories, we are going to use metrics such as – accuracy score, confusion matrix, precision, recall and f1 score etc.

Since the classes in our data sets are balanced, so accuracy is the most used evaluation metric for classification problems with balanced classes.

**Accuracy** – It is one of the common evaluation metrics in classification problems, that is the total number of correct predictions divided by the total number of predictions made for a dataset. Accuracy is useful when the target class is well balanced.

**Confusion Matrix** – A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model. It is a square matrix whose dimensions depend on the number of classes we have in our model. It is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.

**Why to use Confusion Matrix** - Classification accuracy alone can be misleading if we have an unequal number of observations in each class or if we have more than two classes in your dataset. Our data sets contain more than two classes, so Confusion Matrix is the right evaluation metrics.

**Precision** – Precision answers the question of “what proportion of predicted positives are truly positive?” The precision is calculated by dividing the true positives by the sum of true positives and false positives.

**Recall** – Recall answers the question of “what proportion of actual positives are correctly classified?” It is calculated by dividing the number of true positives by the sum of true positives and false negatives.

**F1 score** – Due to their nature, precision and recall are in a trade-off relationship. You may have to optimize one at **the cost of the other. This is where the F1 score comes in. It is calculated by taking the harmonic mean of precision and recall** and ranges from 0 to 1. F1 score is using harmonic mean because harmonic mean has a nice arithmetic property representing a truly balanced mean.

Also, we are going to plot confusion matrix for all classes to see the accuracy of classes with respect to other classes and precision, recall and f1 score are calculated together using Classification Report.

**Model Evaluation Summary**:

|  |  |
| --- | --- |
| **Model Name** | **Accuracy** |
| Naive Bayes Classifier | 0.55 |
| Linear SVM | 0.68 |
| Logistic Regression | 0.70 |
| XGBoost Classifier | 0.70 |

**Challenges and Opportunity:**

The model was trained using past news headlines and categories, the news categories model trained on were limited, if the news with new category needs to be added then model may not predict it accurately. To overcome this challenge, model needs to be continuously trained on new news and categories. This can be done by feeding the news which was already correctly categorized by model or human to improve learning and accuracy. With training on millions news data with several news categories help model learn variety of categories and predict more accurate classes and overall accuracy will be improved over the course of time.

Also, the accuracy of the model can be improved by providing the balanced classes data since we trained this model on balanced classes data but evaluated on dataset with imbalanced classes, despite that it gives good accuracy.

**Recommendation:**

The purpose of this project is to gain insights from news headlines and summary and train model to learn the relationship between news and category. When the model is used on new set of news and categories it should be able to assign each news with appropriate category.

After training data on different ML algorithms and evaluating their accuracy, we recommend the **XGBoost** model to predict news categories which produce highest accuracy and reliable to work in production system. The **XGBoost** model has accuracy of **70%** classifying accurate categories and can be used for news categorization.

As stated above, this model can be beneficial to news editors, owners those who manually categories the news or does not have automated machine learning model to categories news. This model would help them make news categorization process very fast and efficient and they do not need to invest on people who does this task manually or with minimal use of programs.

With use of ML model in production system minimize the human error, categorization process will be faster and accurate, and user would find news into accurate section, this would save human efforts and enhance user count and newspaper brand value.

Also, same model can be used for news recommendation based on past reading history of users and new news feed will be appropriately fed into accurate category, this way user can find their favorite news on their fingertips, and they do not need to search for them.

**Conclusion:**

The problem statement of this project was to build a machine learning model which would help predict the right news category based on news headlines and short description. This process was done manually by reading headlines and summary of news. With this model this process can be automated, and model will predict the news category appropriately.

Our Project, News Categorization, is model to classify news based on news headlines and short summary that would help news editors, owners, and end users.