**Text Categorization using PySpark NLP**

A PROJECT REPORT

***Submitted by***

BL.EN. U4AIE21126 Suryamritha M

BL.EN. U4AIE21128 T.VNL Harshitha

BL.EN. U4AIE21139 Varshini Balaji

**BACHELOR OF TECHNOLOGY**

IN

COMPUTER SCIENCE AND ENGINEERING (Artificial Intelligence)



AMRITA SCHOOL OF COMPUTING, BANGALORE

AMRITA VISHWA VIDYAPEETHAM

BANGALORE 560 035

June – 2023

**ABSTRACT**

The main objective of text classification is to categorize documents into a specific number of predefined categories. We can easily imagine the issue of arranging documents, not by topic, but rather by large assessment, e.g., deciding if the sentiment of a document is whether positive or negative or to which category a product belongs to. While working on a supervised machine learning problem with a defined dataset, there are many classifiers that can be used in text classification. In our model we are using a supervised Machine Learning Algorithm called Logistic Regression to train our model and Techniques like Tokenizer, Stop Word removal, Hashing TF, IDF Vectorizer are used for Computation of the model.

**TABLE OF CONTENTS**

**Page no**

ACKNOWLEDGEMENTS 1

ABSTRACT 2

LIST OF FIGURES

Fig 4.1 Flow chart

CHAPTER 1- INTRODUCTION 4

CHAPTER 2 – LITERATURE SURVEY

5

CHAPTER 3 – REQUIREMENTS AND ANALYSIS 6

3.1 SOFTWARE REQUIREMENTS

CHAPTER 4 – DESIGN AND IMPLEMENTATION 7

CHAPTER 5 – RESULTS 20

CHAPTER 6 – CONCLUSION AND FUTURE ENHANCEMENT 21

REFERENCES 22

**CHAPTER - 1**

**INTRODUCTION**

Text classification, also known as text categorization, is the process of categorizing the text into different organized groups. By using Natural Language Processing (NLP) and suitable Machine Learning model, text classifiers can automatically analyze text and then assign a set of pre-defined tags or categories based on its content. There are different ways of classifying text based on the type of data. They are Single class Text classification, Multi class text Classification and Multi Label Text Classification.

Text classification is a key component of natural language processing. Some of the usages of text classification are spam detection, sentiment analysis, topic labeling, intent detection, Priority detection in customer support, Toxic comment detection.

In our Model we are using Multi Class Text Classification. Multi-class text classification is a text classification task with more than two classes/categories. Each data sample can be classified into one of the classes. However, a data sample cannot belong to more than one class simultaneously.

**CHAPTER -2**

**LITERATURE SURVEY**

[1] Multi-Class Text Classification: Model Comparison and Selection, Proc. of the 3rd International Conference on Electrical, Communication and Computer Engineering (ICECCE)

This study focuses on text classification, aiming to categorize documents into predefined categories. Specifically, the objective is to determine the sentiment of a document as positive or negative. By employing supervised machine learning techniques on a dataset comprising stack overflow questions, answers, and tags, it is observed that machine learning models outperform human baselines. The classifiers considered include Naive Bayes Classifier for multinomial models, Linear Support Vector Machine, Logistic Regression, Word2vec with Logistic Regression, Doc2vec with Logistic Regression, and Bag of Words (BOW) with Keras. The paper provides a comprehensive analysis and comparison of the accuracies achieved by these algorithms.

.

[2]Multi-Class Text Classification Using Machine Learning Models for Online Drug Reviews, 2021 IEEE World AI IoT Congress.

This research analyzes online user reviews in the pharmaceutical field to determine the effectiveness of drugs and their associated side effects. Six supervised machine learning classifiers, including Multinomial Naive Bayes, Multinomial Logistic Regression, Linear Support Vector Classifier (SVC), Decision Trees, Extra Trees, and Random Forests, are used to predict the medical condition based on the reviews. The results show that Linear SVC is the most efficient classifier based on Precision, Recall, F1 score, and training/testing time

**CHAPTER 3**

**REQUIREMENTS AND ANALYSIS**

**3.1 SOFTWARE REQUIREMENTS:**

Python libraries:

* Pyspark- PySpark is a Spark library written in Python to run Python applications using Apache Spark capabilities, using PySpark we can run applications parallelly on the distributed cluster (multiple nodes). In other words, PySpark is a Python API for Apache Spark.
* Pandas- Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data.
* Streamlit- Streamlit is a promising open-source Python library, which enables developers to build attractive user interfaces in no time.
* Pyngrok- pyngrok is a Python wrapper for ngrok that manages its own binary, making ngrok available via a convenient Python API

**CHAPTER – 4**

**SYSTEM DESIGN AND IMPLEMENTATION**

**1.Data Collection**

Collect text data that is representative of the problem you are trying to solve. This data should be labeled with the corresponding classes or categories. There are several ways to collect data for multi-class text classification. Here are some common methods:

* Manual Annotation: In this method, human annotators manually label the text data with the corresponding classes. Annotators read each document and assign the appropriate category or class label based on the content.
* Existing Datasets: There are many publicly available datasets that have already been labeled for specific tasks. These datasets can be found in research repositories, such as UCI Machine Learning Repository, Kaggle, or academic publications.
* Web Scraping: Web scraping involves extracting text data from websites. One can scrape websites that contain documents or articles labeled with categories of interest
* APIs and Data Providers: Some companies or organizations offer APIs or data services that provide access to specific types of text data.

**2. Data Preprocessing:**

Clean and preprocess the text data to prepare it for analysis. This may involve removing irrelevant characters or symbols, converting text to lowercase, removing stop words like "in," "and" and performing stemming or lemmatization to reduce words to their base form.

**3. Feature Extraction:**

Feature extraction is the process of transforming the preprocessed text data into numerical features that can be understood and processed by machine learning algorithms. Few of the methods to extract preprocessed data into numerical features are Bag of Words (BoW), TF-IDF (Term Frequency-Inverse Document Frequency), Word Embeddings and Neural Network based Approaches. The choice of feature extraction technique depends on the nature of the data, the size of the dataset, and the available computational resources. It's recommended to experiment with different techniques to find the most suitable one for your specific task.

**4. Splitting the Data:**

After collecting and preprocessing data, the text data is divided into separate sets i.e., Training and testing to evaluate and assess the performance of the trained model. Training set used to train the machine learning model. The test set is used to evaluate the final model's performance after it has been trained.

**5.Model Selection:**

Choose an appropriate machine learning algorithm for multi-class classification. Algorithms include logistic regression, support vector machines (SVM), random forests, gradient boosting, or neural networks.

Logistic Regression:

Logistic regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set.

A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables

**6.Model training:**

Train the selected model on the training data. During training, the model learns to map the input features (text representations) to the corresponding class labels.

**7.Model Evaluation**

Evaluate the trained model's performance using appropriate evaluation metrics such as accuracy, precision, recall, F1-score, or confusion matrix. This step helps assess how well the model generalizes to unseen data.

**8. Deployment**

Deployment is the method by which you integrate a machine learning model into an existing production environment to make practical decisions based on data.

**FLOW CHART**

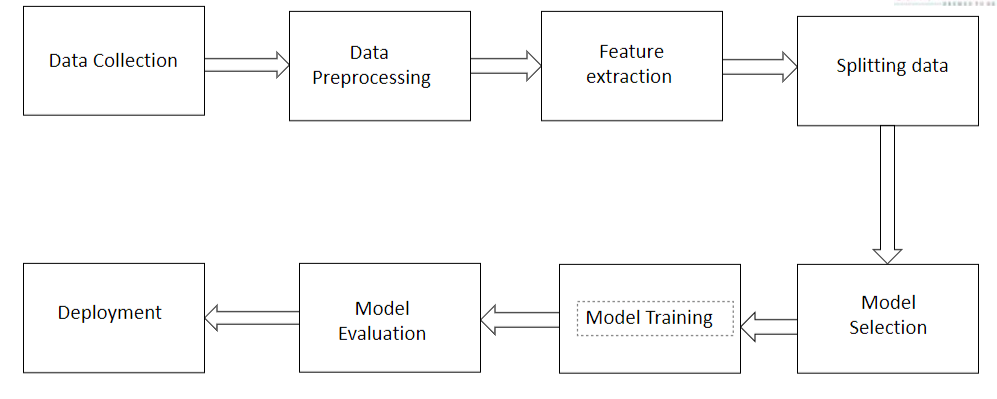
****

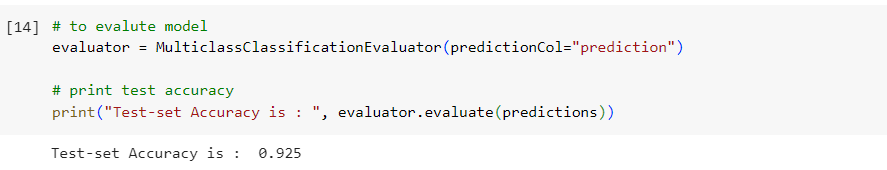
Fig 4.1 System Design

**CHAPTER – 5**

**RESULTS**

Upon training the Logistic Regression model on your news dataset, we have achieved a test accuracy of 92.5%. This accuracy score indicates that the model is performing well in classifying the restaurants.

Accuracy:

Fig 6.1

Web App



Fig 6.2



Fig 6.3



Fig 6.4



Fig 6.5



Fig 6.6

**CHAPTER – 6**

**CONCLUSION AND FUTURE SCOPE**

We developed a text classification model on a news dataset using the Pyspark library. We have used pyspark transformers to vectorize our text. We trained a Logistic Regression model and got 90% of test accuracy and plot confusion matrix. In the end, we put all the preprocessing and modeling steps in a pipeline to run our predictions easily.

**FUTURE SCOPE**

1. Model Optimization: Fine-tune hyperparameters and explore alternative algorithms or ensemble methods to further improve performance.

2. Advanced Feature Engineering: Incorporate techniques like word embeddings, n-grams, and named entity recognition to enhance the representation of text data.

3. Deep Learning Approaches: Explore RNNs and CNNs to leverage their ability to capture complex patterns in text data.

4. Handling Imbalanced Classes: Address class imbalance through resampling techniques or advanced methods like SMOTE or ADASYN.

5. Domain Adaptation: Apply techniques to ensure the model generalizes well to new domains and maintains performance.

By considering these future directions, you can enhance the accuracy, versatility, and scalability of your text classification model.

**REFERENCES**

[1] Multi-Class Text Classification: Model Comparison and Selection, Proc. of the 3rd International Conference on Electrical, Communication and Computer Engineering (ICECCE)

[2]Multi-Class Text Classification Using Machine Learning Models for Online Drug Reviews, 2021 IEEE World AI IoT Congress.