Midsem Report

1st Khushi Patel
Btech Computer Science
(Ahmedabad University)
Ahmedabad, India
khushi.p4@ahduni.edu.in

2nd Devyash Shah
Btech Computer Science
(Ahmedabad University)
Ahmedabad, India
devyash.s@ahduni.edu.in

3rd Aastha Gaudani
Btech Computer Science
(Ahmedabad University)
Ahmedabad, India
aastha.g2@ahduni.edu.in

4th Simran Khoja
Bba Honors
(Ahmedabad University)
Ahmedabad, India
simran.k1@ahduni.edu.in

I. ABSTRACT

Text classification is indeed a crucial task in natural language processing (NLP) that involves categorizing text data into predefined categories or labels. It is widely used in various applications, such as sentiment analysis, spam detection, topic modelling, and content filtering. manually labelling text data is a challenging and time-consuming task, especially when dealing with large datasets. That's why machine learning-based approaches have gained significant attention in recent years, as they can automate the process of text classification and produce more accurate and efficient results. There are various machine learning algorithms and techniques used for text classification, including Naive Bayes, Support Vector Machines (SVMs), Decision Trees, Random Forests, and Neural Networks. These methods use different features and models to learn the patterns and relationships between text data and labels. Text classification is a vital task in NLP that enables efficient data search and analysis. Machine learning-based approaches have made significant progress in automating this task, and further advancements in this field are expected in the coming years.

II. INTRODUCTION

This project aims to perform text category classification on a BBC news dataset using machine learning techniques. The dataset consists of articles published by the British Broadcasting Corporation (BBC) across five categories: business, entertainment, politics, sport, and tech. The goal is to develop a model that can accurately predict the category of a given article based on its text content.

Classifying news articles automatically can have significant practical applications, such as helping news organizations understand their readership better or enabling personalized news recommendation systems. The project will use a variety of machine learning algorithms, including traditional models such as logistic regression and support vector machines.

The main challenges in this project include feature extraction from the text data, dealing with class imbalance (i.e., some categories having fewer samples than others), and optimizing the model's performance in terms of accuracy, precision, and recall. To address these challenges, we will experiment with different text representation techniques, such as bag-of-words and word embeddings, and employ various strategies for addressing the class imbalance, such as oversampling or undersampling.

III. LITERATURE REVIEW

The base paper considered provides an algorithm that includes both the LDA as well as SVM. The algorithm proposed first applies LDA for dimensionality reduction, and applies SVM for classification afterwards. The purpose of using LDA is to separate the samples from different classes so to get such a lower dimensional space. The main function of LDA is to minimize the distance within the class and to maximize the distance between classes considering the assumption that each class follows Guassian probability density function with same covariance. The resulting lower dimensional space is more efficient and discriminative for text representation tha LSI. The Support Vector Machine is applied on the later stage to classify the text. The results observed from the experiments on a benchmark dataset depict the effectiveness of the algorithm that the paper proposes for text classification.

IV. METHODOLOGY

This project aims to develop a machine learning model that accurately predicts the category of news articles from the BBC news dataset. This will involve several techniques, such as data cleaning and preprocessing, feature extraction, and machine learning algorithms. The first step is cleaning the text data using NLTK libraries, then extracting features using TFIDF vectorizer. The model will be trained using SVM and evaluated using an accuracy score. The experiment will be repeated using CountVectorizer and LDA to compare the performance of both techniques. The aim is to demonstrate the effectiveness of various machine learning techniques in text category classification and identify the best approach for accurately classifying news articles from the BBC dataset.

1) Data Collection: The first step was to collect the BBC News dataset. We used the OS and glob modules to import data from each category folder and read the text files.
2) Data Preprocessing: We then performed preprocessing on the raw text data to clean and normalize the text. We removed punctuations, stopwords, and performed lemmatization to extract the root word of each token.

3) Feature Extraction: The next step was to convert the preprocessed text data into numerical features that machine learning algorithms can use. We used the CountVectorizer from Scikit-learn to generate a bag of words representation of the text data.
4) Latent Dirichlet Allocation (LDA): We used LDA, a generative statistical model, to extract topics from the text data. LDA represents each document as a distribution over topics, and each topic as a distribution over words. This enabled us to reduce the dimensionality of the feature space and extract latent topics that can improve classification accuracy.
5) Support Vector Machine (SVM): We then used SVM, a popular linear classification algorithm, to train a model on the LDA topic distribution features and classify the text data into five categories.

V. Implementation

1) Data Preprocessing: The text data was preprocessed by performing the following steps:

   a. Removing punctuation and special characters
   b. Converting all text to lowercase
   c. Removing stop words
   d. Lemmatizing the text

2) CountVectorizer: CountVectorizer is a method for converting text data into numerical features. It counts the frequency of each word in the text data and creates a sparse matrix of the word counts. The resulting matrix can be used as input to a machine-learning algorithm.

3) LatentDirichletAllocation: LatentDirichletAllocation is a topic modelling algorithm that identifies topics in documents. It assumes that each document is a mixture of several topics, and each topic is a mixture of several words. The algorithm determines the topic mixture proportions for each document and the word mixture proportions for each topic.

4) Support Vector Machines (SVM): SVM is a machine learning algorithm that can be used for classification or regression. In the case of text classification, SVM is used to predict the category label for a given input text. The SVM algorithm finds the hyperplane that maximally separates the different classes in the feature space.

5) Accuracy: Accuracy is a measure of how well a machine learning model can predict the correct label for a given input. It is defined as the number of correctly predicted labels divided by the total number of predictions. It can be expressed as a percentage.

   The formula for accuracy:

$$Accuracy = \frac{(Numberofcorrectlypredictedlabels)}{(Totalnumberofpredictions)}$$

6) Prediction of new text inputs: To predict the category label for a new input text, the following steps are performed:

   a. Preprocess the new input text using the same preprocessing steps as the training data.
   b. Convert the preprocessed text into numerical features using CountVectorizer.
   c. Extract the topic distribution features using LatentDirichletAllocation.
   d. Use the trained SVM model to predict the category label for the new input text based on the LDA topic distribution features.

7) TF-IDF: TF-IDF stands for Term Frequency-Inverse Document Frequency. It is a numerical statistic that reflects how important a word is to a document in a collection or corpus of documents. TF-IDF is a measure that combines two factors:
   Term Frequency (TF): the frequency of a term (word) in a document. It measures how often a word appears in a document.
   Inverse Document Frequency (IDF): the inverse of the frequency of the term in the whole collection of documents. It measures how unique or rare a word is across all documents.
   The formula for calculating TF-IDF is as follows:
   TF-IDF = (Term Frequency / Total number of terms in the document) * log(Total number of documents / Number of documents containing the term).

The higher the TF-IDF score of a term in a document, the more important or relevant it is to that document.
The use of TF-IDF and countVectorizer is done in different models. First we have generated a model using TF-IDF for extraction and then training the model using SVM. In the second model we have used countVectorizer to calculate word frequency and we have used LDA to predict the categories of our model, and we have trained the model with SVM.

VI. Results

The accuracy of the model implemented using TF-IDF was 0.09887640449438202, whereas the accuracy of the model implemented with countVectorizer and LDA was 0.8449438202247191.

VII. Conclusion

The difference in the accuracy was because of using LDA as feature extraction in the second model. In the first model prediction were only based on the word frequency, whereas in the second model predictions were based on word frequency

and also with the help of LDA we described our data set as set of categories.

REFERENCES

[1] Support Vector Machine (SVM) algorithm - javatpoint. www.javatpoint.com. (n.d.). Retrieved March 11, 2023, from https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm

[2] Jain, P. (2021, June 1). Basics of countvectorizer. Medium. Retrieved March 11, 2023, from https://towardsdatascience.com/basics-of-countvectorizer-e26677900f9c

[3] Sklearn.svm.SVC. scikit. (n.d.). Retrieved March 11, 2023, from https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

[4] Performance evaluation of latent Dirichlet allocation in text mining ... (n.d.). Retrieved March 11, 2023, from https://ieeexplore.ieee.org/abstract/document/6020066