**Mini Project Report on**



**TEXT CLASSIFICATION**



**Submitted in partial fulfillment of the requirement for the award of the degree of**

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE & ENGINEERING**

**Submitted by:**

**Student Name University Roll No.**

**Ketan Singh Rautela**  **2018881**

***Under the Mentorship of***

**Dr. S S Samant**

**Assistant Professor**



**Department of Computer Science and Engineering**

**Graphic Era (Deemed to be University)**

**Dehradun, Uttarakhand**

**July-2023**



**CANDIDATE’S DECLARATION**

I hereby certify that the work which is being presented in the project report entitled **“Text Classification ”** in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineeringof the Graphic Era (Deemed to be University), Dehradun shall be carried out by the under the mentorship of **Dr. S S Samant, Assistant Professor**, Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun.

Name University Roll no. **signature**

Ketan Singh Rautela 2018881

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **Chapter No.** | **Description** | **Page No.** |
| Chapter 1 | Introduction | **1-2** |
| Chapter 2 | Literature Survey | **3-4** |
| Chapter 3 | Methodology | **5-8** |
| Chapter 4 | Result and Discussion | **9** |
| Chapter 5 | Conclusion and Future Work | **10** |
|  | References | **11** |

**Chapter 1**

**Introduction**

**1.1 Problem Statement-**

Text Classification using Vader Model for Sentimental Analysis of Amazon Food Review

**1.2 Introduction:**

Classification of Texts, also known as Textbook tracking or Textbook classification, is the

process of dividing Textbooks into groups. Using natural language processing (NLP), bibliographers can search for books and prioritize or sort by their content. The potential of informal books is extensive similar to email, social media, websites and many other platforms . However, it can be difficult to extract value from this data unless it is organized in some

way. Doing this used to be a small and efficient process, as manually sorting data takes time and money. Textbooks using NLP have proven to be fast, efficient and can be used to model data.

A bowl of cookies with text

Description automatically generated

Text classification is often the first step in selecting a dataset for further processing, and perhaps the only step in writing a book like a critical review. Entries in parentheses are not intended to provide information in a textbook that is not an order form. One way to teach the classification of books is to use builtin methods to identify methods for describing data and reuse and reuse these features to choose the order to be used for a particular document. Sentimental analysis is one way of doing that.

Areas such as business, product management, education and management already use the

process of extracting and analyzing information from data.

**1.3 Main Libraries Used for the Code are:**

1. **Pandas**- Pandas is an open source Python package widely used for data science/data

analysis and machine learning tasks. It is built on top of another package called Numpy that

provides support for multiple arrays.

2.**Numpy** - We have names for the target array in Python, but their reuse is slow. Numpy's goal is to provide array objects that are 50x faster than Python lists. Array objects in Numpy are

called ndarrays and provide many support functions that make working with ndarrays very

easy. Arrays are constantly used in data smart works where speed and money really matter.  
  
3.**NLTK** - NLTK stands for Natural Language Toolkit. It is basically a powerful Python

library that provides tools and resources for manipulating human language data in natural

language processing (NLP). NLTK was developed by the Natural Language Processing Group at the University of Pennsylvania. NLTK provides many functions, including text processing,

token generation, rooting, part of speech tagging, and name association.  
  
4.**Seaborn** - Seaborn is a Matplotlib based data visualization library for Python. It provides maximum links to create beautiful and informative reviews. Seaborn is designed to work seamlessly with the pandas data framework, making it a popular choice for data visualization in data analytics and exploratory data analysis (EDA).

**1.4 Applications:**

1.Sentimental Analysis.

2.Spam Detection.

3.Language Identification.

4.IntentClassification.

**Chapter 2**

**Literature Survey**

This article represents a research contribution that has already been made in education. The use of text classification has become effective and accurate in many research papers. Some of

them are as following-

As the number of data increases, so does the computational complexity (Stas, Juhar & Hladek 2014). When it comes to data, machine learning is often considered a branch of statistics. He uses advanced models to make assumptions based on his own experience (Du, 2017; Ranjan & Prasad, 2017). However, statistical and mechanical approaches are considered inadequate, so a combination is usually preferred (Srivastava, 2015).

“ A Comparison of Event Models for Naive Bayes Text Bracket" by Andrew McCallum and Kamal Nigam (1998).This influential paper introduced the use of Naive Bayes classifiers for textbook bracket and compared different event models( Bag of Words, Bigrams, Trigrams) on a range of datasets. It laid the foundation for textbook bracket using statistical machine literacy ways.

"AutoText Classification Using Neural Networks", Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner (1998): This article examines the application of neural networks, particularly multilayer perceptrons (MLPs), to the text classification task. It demonstrated the potential of neural networks in processing complex data and paved the way

for deep learning in NLP.

"Support Vector Machines for Text Classification" by Thorsten Joachims (1998) This article demonstrates the effectiveness of support vector machines (SVMs) for the text classification task. The results show that SVM outperforms traditional methods such as Naive Bayes on rich and colorful data and is gaining popularity as an important scaffolding method.

"Convolutional Neural Networks for Judicial Bracket" By Yoon Kim (2014) This article presents the performance of Convolutional Neural Networks (CNN) for judgment location manual

classification. CNNs were originally designed for image recognition, but they have also proven effective at classifying books.

"Attention All You Need" (2017) seminal article by Vaswani Etal, introduces the

Transformer luminaire, which attaches great importance to the continuous reuse of materials. Mills has revolutionized NLP activities, including the scaffolding book, and models such as BERT (Mills Represented Bidirectional Encoder) have achieved case results.

"XLNet: Generalized Autoregressive Pre-Training for Understanding Language", Yang et al. (2019): XLNet is an extension of the BERT model that uses permutation training to capture bidirectional content while avoiding the limitations of traditional masking models. Improves BERT performance on many tasks, including text classification.

"ULMFiT: Universal Language Model FineTuning for Text Classification", Jeremy Howard and Sebastian Ruder (2018): This article presents ULMFiT, a learning transformation for text classification using predefined language models. ULMFiT shows significant performance improvements with minimal data loss.

A diagram of a diagram of a system

Description automatically generated

This case study highlights some of the key advances in text classification techniques, from statistical methods to today's deep learning models.

**Chapter 3**

**Methodology**

Text classification involves many steps and techniques to accurately process, represent and

classify data.

**3.1 Pre-processing the Dataset:**

The first step is to remove all unnecessary data from the file transfer. We will undo the beginning and end of the dataset. This is information that is not used by us. Data preprocessing is

the process of converting raw data into an understandable format. This is also an important step in data discovery because we don't have raw data.  
Data quality should be checked before using machine learning or data mining algorithms.

Finally, we'll make sure we only have private messages. This will help the model work better to avoid confusion due to word repetition.

**3.2 Import Libraries and Dataset:**

**We import some of the important libraries of python like pandas, numpy , NLTK, seaborn and matplotlib for the working of our project-**

A computer screen shot of a black background

Description automatically generated

Importing of Dataset plays an important role it is second step after importing of libraries as all the operations will be done on these dataset. Importing of dataset will be done by csv library by passing the path of that file.

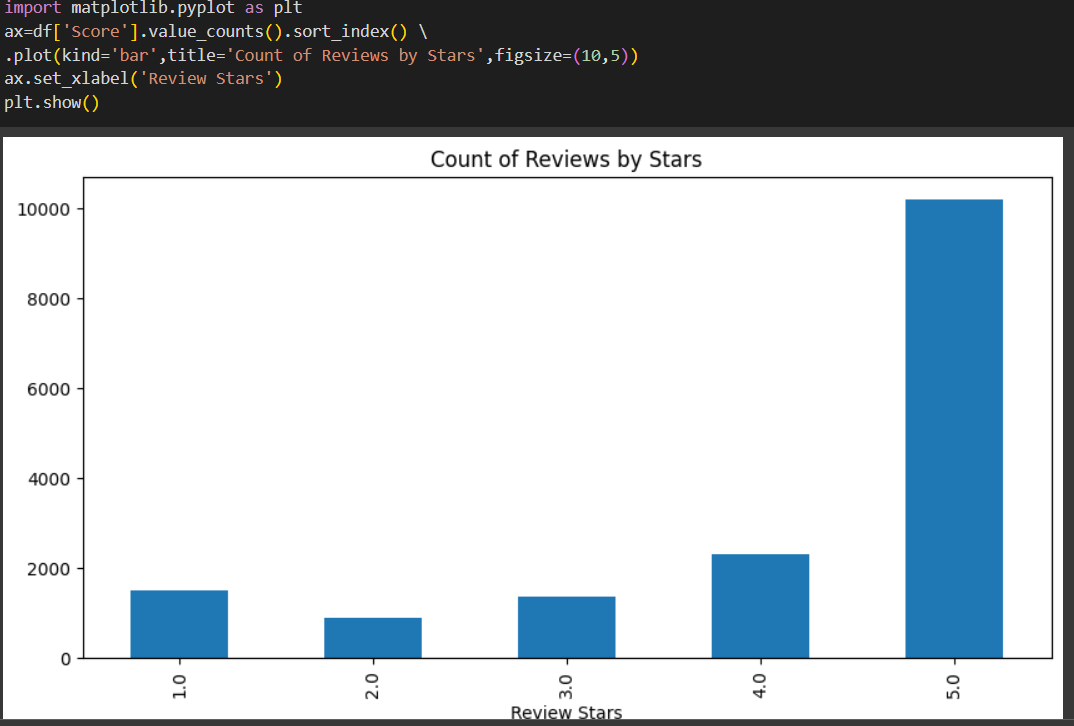
****

**3.3 Feature Extraction and Representation:**

a. Convert the preprocessed text data into the numerical feature vectors that can be processed by machine learning algorithms.

b. Few of the techniques include Bag-of-Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), and word embeddings (e.g., Word2Vec, GloVe).

**3.4 Visualizing the Model:**

****

**3.4 Model Training:**

1. Split and separates the dataset into training and validation sets to train the model and evaluate its performance.
2. Enter the unique number vectors and corresponding labels into the models selected for training.
3. Optimize hyperparameters by methods such as grid search or random search.

**3.5 Tokenization:**

Tokenization is an important step in text classification, where a document is divided into groups called tokens. Symbols can be words, subwords, symbols or phrases, depending on the symbolization strategy used. Symbolization is the first step in transforming raw text into a format that machine learning models can process.  
Different symbolization methods are used in text classification, and the choice of symbolization method may affect the performance of the model. 

When tokenized, data files are typically processed to perform operations such as word extraction, root extraction, and feature extraction (such as word bag or TF:IDF) to generate numerical representations for machine learning models. The processed data is used to train and evaluate text classification models by enabling them to predict appropriate classes or categories for new, unseen text.

**3.6 Import Sentimental Analyzer:**

You can use many libraries and tools in Python to make assumptions for text classification. The Natural Language Toolkit (NLTK) library, which provides logic tests, is a popular choice. Using

NLTK - To deploy the logic tests before using the  
NLTK library, we need to import that library and then deploy the logic tests. 

**A black background with white text

Description automatically generated**

**3.7 Import Vader Model :**

To use the VADER (ValenceAware Dictionary and Sentiment Reasoner) model for sentiment analysis in text classification, you need to install the vaderSentiment library. VADER is a dictionary and sentiment managementbased tool designed for sentiment analysis social media articles and short articles.

**A computer screen shot of a computer code

Description automatically generated**

**3.8 Model Testing and Deployment:**

a. Once satisfied with the model's performance, test it on a separate test dataset to assess its generalization capabilities.

b. Deploy the trained model to make predictions on new, unseen text data.

They useded predicting next character high-

lighter(PNCH) for Indian language it was more of

text correction and less about next word prediction but

was quite good to understand. The Method called hit and

miss but accuracy is less and the model was not eﬃcient

for this kind of problem statement.

Authors have used LSTM for next word prediction

for Assamese language. They have stored transcripted

language according to International Phonetic Association

(IPA) chart and fed to their model . They created a

model for physically challenged people. This model uses

Unigram, Bigram, Trigram, based Approach for next

Authors have used LSTM for next word prediction

for Assamese language. They have stored transcripted

language according to International Phonetic Association

(IPA) chaedict the

word but accuracy was around 30-40 percentage

**3.9 Performance Analysis and Iteration:**

a. Analyze the model's performance on real-world data and monitor its behavior in production.

b. Iteratively improve the model based on feedback and observations from users or stakeholders.

A diagram of a diagram

Description automatically generated

**Chapter 4**

**Result and Discussion**

The model is used for sentiment analysis of Amazon's food inspection data and some important Python libraries such as Pandas, Numpy, seaborn, matplotlib and two models used for data analysis of data, Vader Model and Sentimental Analyzer. The dataset includes consumer

reviews of food products. Using models, we make statistical analysis of them in the form of

graphs.

**A graph of different colored bars

Description automatically generated**

These are the results we get with the help of Vader model and Sentimental Analyzer.

These are the results we obtained with the help of Vader model and emotional evaluation.  
In general, the manual binding method forms the basis of all research methods. Because they provide the structure of the raw data. Some of the major issues in manual system are removing tokens first, stopping message, pointing to remove unknown content, storage, access, parameter estimation data instability overfitting etc.  
During data mining, it is difficult to extract deep meaning or content from data.  
Semantic technology faces more problems in language processing situations, especially

automation. This is mainly due to the problem of ambiguity in natural language. Emails,

abbreviations, SMS codes, etc. in text files. The presence of heterogeneous components such as adding the competition to existing data mining tools, as each requires a different algorithm for ranking. Reducing the amount of data remaining in the database is another important factor for text analysis. While this is a data security concern, the need for text analytics is security.

**Chapter 5**

**Conclusion and Future Work**

We created this project by analyzing sentiment using the Vader model of the Amazon food review dataset, which shows all the stats of reviews such as positive, negative, and neutral. This will be done by the Sentiment Analyzer model and the Vader model. Both of these models

are useful for analyzing emotions. This project will help in many areas such as spam

detection, target classification, search terms.  
shows a way to create a general structure that combines the decision structure and the

hierarchical bracket structure into a single structure.

Therefore, the biases of the judgment model and the disadvantages of the hierarchical

classifier can be developed simultaneously at each educational level. However, according to the literature, semiscriptive classifications have become important in writing due to their efficient classification. It reduces time costs. Some of the key issues include improving

performance, handling large taxonomy, selecting features, data regions, and data

inconsistencies.

Naive Bayes can be improved by increasing the prior probability and modeling iteratively.  
For example, models developed using a modified Naive Bayes algorithm were used to develop flood analysis. The experiments in this article show that test results can be better obtained before the event, which will help improve the later prediction of the event and reduce the uncertainty in frequency calculation. But its performance in the distribution of letters needs to be further improved.

Based on data analysis, data collection, feature extraction, and classification algorithms are the three most important areas to improve  .

**References**

1. Jordan, Michael I., and Tom M. Mitchell. "Machine learning: Trends, perspectives, and prospects." Science 349, no. 6245 (2015): 255-260.
2. Sahoo, Abhaya Kumar, Chittaranjan Pradhan, and Himansu Das. "Performance evaluation of different machine learning methods and deep-learning based convolutional neural network for health decision making." In Nature inspired computing for data science, pp. 201- 212. Springer, Cham, 2020.
3. Prajapati, Gend Lal, and Rekha Saha. "REEDS: Relevance and enhanced entropy based Dempster Shafer approach for next word prediction using language model." Journal of Computational Science 35 (2019): 1-11.
4. Ambulgekar, Sourabh, Sanket Malewadikar, Raju Garande, and Bharti Joshi. "Next Words Prediction Using Recurrent NeuralNetworks." In ITM Web of Conferences, vol. 40, p. 03034. EDP Sciences, 2021.
5. Stremmel, Joel, and Arjun Singh. "Pretraining federated text models for next word prediction." In Future of Information and Communication Conference, pp. 477-488. Springer, Cham, 2021.
6. Xiaoyun, Qu, Kang Xiaoning, Zhang Chao, Jiang Shuai, and Ma Xiuda. "Short-term prediction of wind power based on deep long short-term memory." In 2016 IEEE PES AsiaPacific Power and EnergyEngineering Conference (APPEEC), pp. 1148-1152. IEEE, 2016.
7. 7.Vargiu, Eloisa, and Mirko Urru. "Exploiting web scraping in a collaborative filtering-based approach to web advertising." Artif. Intell. Res. 2, no. 1 (2013): 44-4.