BEYKOZ ÜNİVERSİTESİ

GRADUATEPROGRAMSINSTITUTE

URDU NEWS CATEGORIZATION USING MACHINE LEARNING APPROACHES

Muhammad Talha SATTI

MASTER'STHESIS

İstanbul - 2023

URDU NEWS CATEGORIZATION USING MACHINE LEARNING APPROACHES

Muhammad Talha SATTI

MASTER'S THESIS

COMPUTER ENGINEERINGDEPARTMENT

COMPUTER ENGINEERINGMASTER'SPROGRAMWITHTHESIS

MASTER'STHESISADVISOR

Asst. Prof. Dr. Özlem Feyza Erkan

MASTER'STHESISJURYMEMBERS

Dr. BurcinKulahcioglu

Dr. Selcuk Kiran

PREFACE

In the name of Allah, who is the most kind and forgiving. First of all, I would like to thank my supervisor, Dr. Ö. Feyza Erkan, for being such a great guide during my Master's thesis. Dr. Ö. Feyza Erkan, your constant support, lots of patience, reviving motivation, vast knowledge, and logical thinking helped me while I was researching and writing this thesis. I also like that you asked me tough questions, which pushed me to look at my research from different angles. I am sure I wouldn't have been able to do this job without your kindness and hard work, which I really appreciate. Besides my supervisor, I am also grateful to my parents for their continuous prayers for my success.

CONTENTS

ABSTRACT

URDU NEWS CATEGORIZATION USING MACHINE LEARNING APPROACHES

Rapid technological developments have changed the way of presenting news content in media. The last three decades have witnessed the emergence of digital media platforms that offer a broad variety of news content ranging from business to, economics, from Science&Technology to sports. Due to massive amount of data produced, the need for automatically categorizing them has arisen. Motivated by this, we have addressed the problem of text categorization of Urdu news by using four machine learning approaches namely (Naïve Bayes, Support Vector Machines, Logistic Regression and Decision tree). In the first step we have collected the data which contains 4000 news belonging to four different categories (Sports, Business-&-Economics, Science-&-Technology and Entertainment). The data is in the raw format so before setting up the machine learning model, we applied pre-processing techniques like tokenization, removing the stop words and lemmatization. Then, the features are extracted Bag of Words and Term Frequency-Inverse Document Frequency methods. In the last step, we evaluate the performance of machine learning algorithms utilizing accuracy, precision, recall, and F1-score metrics. In future work, deep learning models can be can be utilized provided that more data is collected and a system can be developed which performs real-time categorization.

Keywords: Text Mining, News Classification, Feature Extraction, Machine Learning

ÖZET

MAKİNE ÖĞRENMESİ YAKLAŞIMLARI KULLANARAK URDUCA HABERLERİN KATEGORİZASYONU

Hızlıteknolojikgelişmelermedyadahaberiçeriğininsunuluşbiçiminideğiştirmiştir. Son otuzyıl, işletmedenekonomiye, bilimveteknolojidensporakadargenişbiryelpazedehaberiçeriğisunandijitalmedyaplatformlarınınortayaçıkışınatanıkoldu. Üretilenbüyükmiktardaverinedeniyle, bunlarıotomatikolarakkategorizeetmeihtiyacıortayaçıkmıştır. Bu gelişmelereparalelolarak, Urducahaberlerinmetinkategorizasyonuproblemidörtmakineöğrenmeyaklaşımıkullanarak (Naïve Bayes, DestekVektörMakineleri, LojistikRegresyonveKararağacı) elealınmıştır. İlk aşamada, dörtfarklıkategoriye (Spor, İş&Ekonomi, Bilim&TeknolojiveEğlence) ait 4000 habermetniiçerenverilertoplanarakbirverikümesioluşturulmuştur. Toplananverilerişlenmemişolduğundanmakineöğrenimimodelikurulmadanönce, dizgeparçalama, durdurmasözcüklerininkaldırılmasıvebaşkelimeyibulmagibiönişlemeteknikleriuygulanmıştır. Dahasonraöznitelikler, KelimeTorbasıveTerimSıklığı-TersDokümanSıklığıyöntemleriyleçıkartılmıştır. Son aşamada, doğruluk, kesinlik, geriçağırmave F1 puanımetriklerikullanılarakmakineöğrenimialgoritmalarınınperformansıdeğerlendirilmiştir. İlerideyapılacakçalışmalarda, dahafazlaveritoplanmasıilederinöğrenmemodellerindenyararlanılabilirvegerçekzamanlıkategorileştirmeyapanbirsistemgeliştirilebilir.

AnahtarKelimeler: MetinMadenciliği, Haber Sınıflandırma, ÖzellikÇıkarma, MakineÖğrenmesi

ABBREVIATIONS

NLTK : Natural Language Toolkit

NLP : Natural Language Processing

SVM : Support Vector Machine

BOW : Bag of Words

KNN : K-nearest Neighbour

TF-IDF: Term Frequency Inverse Document Frequency

LISTOFFIGURES

LISTOFTABLES

1 INTRODUCTION

Urdu which is a prominent language in Indo-Pakistanis spoken by more than 100 million people in different parts of the world. Urdu is one of the India's twenty-three official languages, as well as the national language of Pakistan.

It is the twenty-first most common language spoken throughout the globe. Because more people are using the internet, the amount of data stored electronically is rapidly expanding. In recent years researchers have developed a growing interest to extract information from enormous data volumes in a quick and efficient manner. Text categorization is crucial and has several applications, including recognizing the styles of the text, filtering the news stories according to the user's preferences, and assessing whether or not an e-mail is spam. When we want to give each item a different category label, we could find that text classification is helpful for article labelling since it allows us to group articles into particular categories. Text categorization in Urdu languagerequires a greater research effort.

Text mining is used to a large extent in a broad variety of fields, and its contributions are often recognized and praised. Text mining is regarded as the most effective method for trying to extract information from unstructured text by industry professionals. The increasing volume of written content that requires accurate text categorization has led to a rise in the quantity of research conducted on text mining.(Devi, Bai and Reddy, 2020) One of the topics that Natural Language Processing often discusses is that of classifying different types of text. It is possible that it is one of the most popular topics to investigate in order to get a deeper comprehension of the concepts that underlie machine learning and natural language processing. There are a

number of algorithms that can classify text into its respective categories. It is well known that the typical natural language processing algorithms focus mostly on words in order to choose specified groups for certain texts or text documents.(Hassan, 2016)A computer is provided with access to data and given the instruction to acquire new knowledge about a topic on its own while participating in machine learning. There are three different ways that data may be collected. Discovering natural categorizations in data is the focus of unsupervised learning, which is a kind of machine learning. The computer is oblivious to the existence of classes and is unable to determine which category an item falls into. In supervised learning, predefined classes are used, and the algorithm is capable of determining which categories a piece of data should be placed in. Text categorization refers to the process of assigning textual material into a category. The use of machine learning models achieves this process automatically. It is possible to assign a document to one category, many categories, or none at all.

Textual resources come from many different fields, such as journalism, medicine, economics, academia, and more. It has been shown in the research that being able to design rules to accomplish text classification and to arrange the classified text into text categories is very valuable for categorization of news and knowledge discovery. Although text classification has been used to categories news in different languages, relatively few works reported on categorizing Urdu news,Therefore, the objective of this thesis to investigate text categorization on Urdu news.

1.1 PROBLEM STATEMENT

Most of the studies on the area of text categorization focus on the analysis of English text. However, the proposed models do not apply to the dialects of South Asian languages like Urdu and Persian since they are not used in those contexts. Those languages have their own unique morphology, alphabet, and rules for grammatical construction. When it comes to organization, the fact that there has been few natural language processing (NLP) works done on Urdu is a big cause for concern. Due to the limited number of research done on the Urdu language, there is no such thing as a Lexicon. Urdu is a language that may be spoken in either direction, and it uses an alphabet that is based on Arabic. In Roman Urdu, it is common practice to write English words using Latin characters since this is how the English language was first written. As a consequence of this, we adopted supervised learning techniques for the categorization

of Urdu news. When compared to the English language, it is noted that Urdu is considerably different from English in terms of both its morphology and its script. Because of this, the need for a model that can assess the content of Urdu news articles has increased. Tweets on current events in Urdu were analyzed to compile the data for this study.

In this thesis, we employed supervised Machine learning techniques for news categorization on Urdu language. As a first step, the algorithms are used to preprocess the data set in order to filter it and get rid of noise in the data set. After preprocessing step, the feature selection algorithm is used and then machine algorithm is applied on the dataset and in last result and analysis were compared. Figure 1.1 shows the most general way to look at the derived approach that is used for news categorization.

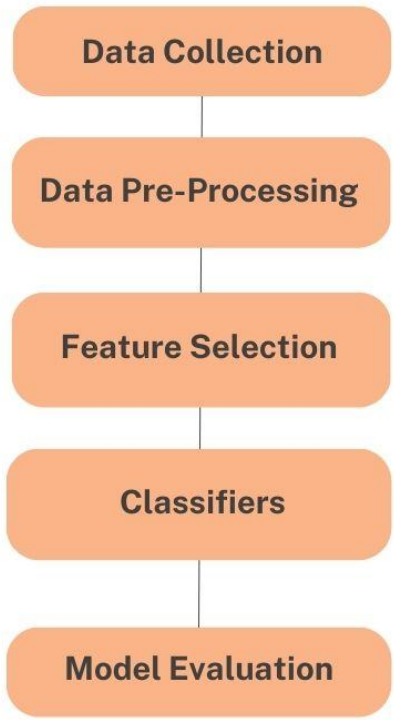

Figure 1.1Main Stages of News Categorization

In this work, we address the following research questions:

RQ1: In Urdu text categorization, which text pre-processing techniques are available?

RQ2: Which techniques have been used for classification of Urdu text and what are the proposed methods for efficient classification of Urdu News?

RQ3: How do different machine learning models perform?

1.2 MOTIVATION

Due to increasing textual information generated in Internet the world wide web, the process of text categorization has attracted several researchers. Text categorization is a sub domain of text classification which is an NLP task. To begin, the difficulty of completing text categorization is an issue for the whole scientific community as a whole. As a direct consequence of this, categorization has developed into a significant area of focus for research within the domain of Natural Language Processing (NLP). While there is extensive research addressing text categorization task in English, only a limited number of works has studied the problem in different languages such as Urdu, Bengali, Hindi etc. Motivated by this we investigate text categorization problem on Urdu language. This problem is quite challenging since Urdu is morphologically rich.

1.3 OUTLINE

The rest of the thesis is organized as follows:

Chapter 2presents a comprehensive review on text classification process also different techniques and approaches have been discussed.

Chapter 3describes the methodology for text categorization which is a sub-domain of text classification problem. The steps of data preprocessing are also explained in detail.

Chapter 4discusses the experimental results and analyzes the performance of the models proposed.

Chapter 5summarizes the achieved results and suggests some future directions for the text classification on Urdu language.

2 LITERATURE SURVEY

In the last few decades, there has been a lot of progress in the field of NLP and its subdomains. One of the problems considered is the text categorization.

Li, Chen, Wen, Jin and Shi (2015) presented a method for forecasting the movement of Chinese stock prices using the text categorization. The representation of the text, the selection of relevant features, and the categorization of the text are the three stages that comprise the model. Both KNN and SVM are used as classification algorithms when dealing with text. One thousand different Poly Real Estate news stories have been collected for experimental setup. The SVM model, which makes use of processes and methods, generates outcomes that are more accurate, with an accuracy rate of 83 percent.

Duwairi, Al-Refai and Khasawneh(2009) evaluated the effectiveness of three classifiers in the Arabic language: the Naive Bayes, theKNN, and the distance-based classifier. The data contains ten main categories with 1000 records in total. Data cleaning has been applied for removing formatting tags, stop words, and punctuation marks. Thenthe texts were categorized using the aforementioned algorithm. According to the findings, the Naive Bayes classifier performed much better than the other two classifiers.

According to Purohit, Atre, Jaswani and Asawara(2015),text categorization refers to the act of arranging texts into distinct categories on the basis of the information

contained within them. Text understanding systems, which change the text in some manner, such as by generating summaries, answering queries, or extracting data, require text categorization in order to function properly.

Text retrieval systems, which retrieve texts in response to a user query, require the categorization as well.The authors used Porter Stemmer and Tokenizer in order to extract keywords from inside this system. The word set may be generated from the derived keywords using both the Association Rule and the Apriori method. The likelihood of the word set is estimated with the help of a naive Bayes classifier. The newly created abstract that the user inputs is then assigned to one of the several classes that are available. The accuracy of the system is considered to be acceptable. It requires a less amount of training data in comparison to other categorization algorithms.

Using Arabic text, the effectiveness of two well-known text classification algorithms, namely SVM and C5.0, was analysed and compared (Al-Harbi, Almuhareb, Al-Thubaity, Khorsheed and Al-Rajeh, 2008). As a first step, the text taken from the papers was transformed into a vector of characteristics. The use of Chi-Squared statistics contributed to a reduction in the amount of space that is needed for vector input. On the basis of the document frequency, the Chi-Squared test is carried out, and the leading 30 class-related features are selected. After that, the data are separated into a training data set consisting of seventy percent and a test data set consisting of thirty percent. During the process of text classification, the algorithms that just went over were used, taking into account the words that were selected. The performance of C5.0 (with an accuracy of 78.42 percent) is better to that of SVM in each and every category of the data (with an accuracy of 68.65 percent).

Maneka and Radha(2013) This work presents a technique for categorizing text, with the extraction of keywords. Both TF-IDF and WordNet are used in the process of keyword extraction. WordNet, which is a lexical database for English words, determines the degree of similarity that exists between the words that are provided. TF-IDF is responsible for providing the words that have the potential to be utilized as keywords, and WordNet is responsible for providing these words. The categorization of the text makes use of the techniques of Naive Bayes, decision trees, and KNN, respectively. In order to assess the algorithms, the 10 folds cross-validation approach was used, and

both training data and test data were employed. The findings reveal that the Naive Bayes method achieved the most accurate result, with an error of 0.3 Root Mean Square.

The term "text categorization" describes the process of classifying texts into various groups according to the data they contain. It is the process of automatically classifying materials written in natural language into a set of categories in advance. (Kamruzzaman, Haider and Hasan, 2010) presented a unique approach for text classification that is based on data mining and requires less training materials than traditional methods. Instead of using individual words, word relations or association rules that are formed from these words are used in the process of extracting a feature set from pre-classified text documents. The resulting features are then put through a Naive Bayes classifier, and ultimately, a single Genetic Algorithm is added for final classification. The results of the experiments showed that the proposed system performs quite well as a text classifier.

In work of (Han, Karypis and Kumar, 2001)Theyperformed text classification using WAKNN,which stands for weighted k-nearest neighbour. When utilizing WAKNN, the complete training set is converted into a matrix, and the contents of each cell indicate the frequency with which a given word occurs in a given text. This concept is referred to as word frequency (TF). Since this matrix has been "normalized," all of its values lies within the range of 0 to 1. In order to determine the degree of similarity between the documents, they used the cosine similarity. This requires a document vector as well as a weights vector to be provided. These weight vectors are altered in order to get the greatest possible outcomes. The project makes use of a variety of algorithms, including WAKNN, K-nearest neighbour, and C4.5, among others. The results of these tests indicate that WAKNN is the approach that yields the most accurate results when compared to others.

Ivana and Masayu, (2016) established a hierarchical structure for the organization of the Indonesian news articles. This study investigated whether or not the outputs of applying hierarchical multi-label classification to Indonesian news articles provide acceptable results. The top-down strategy, the balanced strategy, and the global strategy

are the three primary types of approaches that were considered. When dealing with multi-label classification problem, most of the research relied on binary relevance and calibrated label ranking. The data was gathered from the website of a news organization and arranged in a hierarchical manner. There are ten primary categories into which the news articles may be placed, and each of these categories has its own subcategories. For classification, Naive Bayes, and support vector machine methods were used.

Al-Anzi and AbuZeina, (2018) The authors presented hierarchical text classification of Arabic texts for the first time in the literature.Nevertheless, a number of studies suggest that research on the categorization of flat Arabic text had been done. The first-order Markov chain is applied for categorizationof Arabic texts in a hierarchical fashion. When a text categorization task is carried out using the Markov chain method, a transition probability matrix is generated for each category. This matrix represents the likelihood of transitioning between categories.

A scoring procedure is applied, one training set category at a time, for each individual document. Thestudy makes use of data obtained from the Alqabas newspaper. The documents are arranged in a hierarchical fashion with three levels. They start by doing some preliminary processing on the data. Performing actions such as these helps get rid of data that isnot essential. The pre-processing procedure makes use of the TF IDF file format. After that, Markov chain matrices are used in order to train the data. A number is assigned to each character in the Arabic script which is referred to as mapping. A probability matrix is constructed for each category individually. The authors tested their proposed method using LSI and reported their findings. According to the findings, performance at the top layer is rather solid, but performance at lower levels continues to worsen.

The purpose of the study was to identify the current state of knowledge about the classification of text in order to facilitate the process of determining the ensuing various stages (Bhumika, 2013). To accomplish this objective, the steps involved in the Text Classification process have been laid out. These steps include document collecting, pre-processing, indexing, feature extraction, classification, and accuracy evaluation.

Chy, Seddiqui and Das (2014) developed a model using the Naive Bayes classifier for the purpose of classifying Bengali (language of Bangladesh) news into a total of 34 unique categories. After a dataset was gathered via the use of a custom-built web crawler from an online news aggregator, Naive Bayes was used for the categorization of the news. However, the recall-precision graph that they presented was off by more than twenty percent and failed to accurately identify any of the news categories.

Hossain, Sarkar and Rahman (2020) The authors created a system that could properly categories Bangla news by using a range of machine learning techniques and deep learning models. The findings of a thesis project that employed CNN and Bi-LSTM to categories 12 categories of news items using a dataset gathered from an open-source Bengali corpus showed enhanced accuracy. The dataset was collected from an online Bengali dictionary. In order to classify the articles in the news, a total of four separate Machine Learning algorithms and Deep Learning models (Bi-LSTM and CNN) were used, and the results were analyzed for their degree of accuracy.

3 METHODOLOGY

This section contains the techniques that are used for Urdu text categorization. The stages consist of collection of datasets, dataset pre-processing, feature extraction and model development, respectively.

3.1 DATASET

Corpus building is an essential aspect in the text categorization problem. Apart from English when it comes to the Urdu language, there are limited number of resources that can be used for categorization purposes. In most cases, information is gathered through social networks and internet forums, as well as from newspapers. For the collection of text data,the use of online resources, such as web blogs, social media sites, online news or electronic journals, etc., are the most appropriate and straightforward sources.

The dataused for this thesis is acquired in raw format from the websites of Pakistani conventional news channels, such as ARY News, Geo News, Express News, Dawn News, and Jang News, amongst others. The duration period of our dataset is from year 2015-2020.

The data is then organized and tagged to have three columns, headlines, news text and the category. Headlines includes the main title of the news. News text column consists of the whole text of the news and the category column includes the name of the category of news. Data set is distributed into 4 distinct categories as Business_&_Economics, Entertainment, Sports and Science_&_Technology. The dataset is balanced whereeach category has 1000 text-instance as shown in Table3.1.

Table 3.1 Statistics of Dataset

Category	Value
Business & Economics	1000
Entertainment	1000
Sports	1000
Science & Technology	1000

3.2 DATA PREPROCESSING

It is critical to do preprocessing on text data in order to convert it into a format that is simpler to understand and to improve the training performance of the classification model. In the first step,the text was cleaned up by converting all the uppercase words into the lower-case and omitting special characters such as punctuation, tabs, line breaks, and so on.

In order for the machine learning models to be able to properly learn from the material, the language had to first be parsed in a way that it could be understood by the algorithm. Tokenization refers to the process of transforming the text into a vector of tokens, which may be defined as breaking down the strings into words. In this study, individual words were considered to be "tokens," and the representation of each token inside the dictionary was an integer that is connected to the index of that particular token.

Along with this, stop words have been removed so that the corpus may be cleaned up even more. Stop words are words that aid in the formation of the structure of a sentence but do not provide any further meaning to the message that the sentence is aiming to express(for example, **خو**,**گی**, **کرو**). After removing the stop words, we apply the lemmatization on the data set. Lemmatization is the process of getting the lemma (root word) of the words. Python was chosen as the language to be used for the implementation of the tokenization algorithm to achieve filtering and elimination of noise from the data, removing the stop words and for the lemmatization.

Steps involved in the pre-processing stage is given in the Figure3.1.

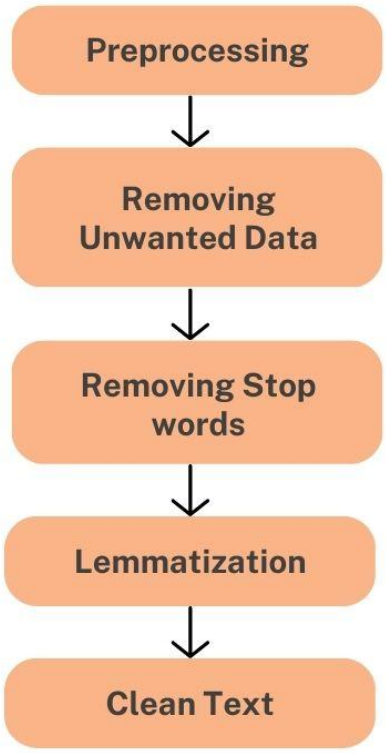

Figure 3.1 Flow Chart of Preprocessing

Algorithm which is used to remove the unwanted data from the Urdu news text is depicted in Table3.2.

Table 3.2 Algorithm for pre-processing the Urdu Text

Input: Urdu News Text **Output:** Returning the clean text
Utilizing the list to save the output result • Removing the links and url • Removing all the @username • Filtering all the special characters;%()|+&=*%.,!?:#$@ Returning **Preprocessed Text**

3.2.1 Tokenization

The process of generating meaningful tokens from raw textual material is referred to as tokenization (Yogish. Manjunath and Hegadi, 2018). Tokenization is the process of recognizing tokens that are included inside documents. This step of text categorization not only increases storage but also fulfills the information requirements of a user while using less search space. Vijayarani and Janani (2016)evaluated the effectiveness of seven different open-source tokenization technologies. Tokenizing an archive may be done using one of the different tools. There are a variety of tools at your disposal, including the Nlpdotnet Tokenizer, the Mila Tokenizer, the Text Blob Word-Tokenizer, the MBSP Word-Tokenizer, the PatternWord-Tokenizer. The most efficient method of tokenization is shown to be Word Tokenization with Python NLTK. The maximum number of characters that may be stored in this application at one time is 50,000. Initially, the text is broken up into phrases by utilizing the WordPunct Tokenizer to do the tokenization. The idea of the algorithm which is used to tokenize the news using WordPunct is given in Table 3.3.

Table 3.3 Algorithm to Tokenize Text

Input: Preprocessed Text **Output:** Tokenize Text
The words are Tokenize using WordPunct Tokenizer Algorithm and appending the Tokenize Sentences **Returning Tokenized Text**

An example output obtained after tokenizing text into words is presented in Table 3.4.

Table 3.4 Example on Word Tokenization

	Text Data
Input News Text	عالمی بینک عسکریت پسندی متاثرہ خاندانوں معاونت گا
Word Tokenization	'عالمی', 'بینک', 'عسکریت', 'پسندی', 'متاثرہ', 'خاندان', 'معاونت', 'گانا

3.2.2Stop Words

The words that are used to finish sentences are referred to as stop words(Khan, Amjad, Ashraf and Chang, 2021). In Urdu, some of the most often used terms are گیandکرو (do). These terms have been eliminated from our corpus. Nevertheless, because of the morphological nature of the Urdu language and the limited resources available, it is difficult to get rid of stop words automatically. These include helping verbs, articles, and prepositions. Using the website ranks[1], we were able to compile the stop words data setof Urdu. The list consists of 553 stop words. Some common stop words are shown in Table 3.5.

Table 3.5Someexamples of Stop words

گی	کرو	خو	کرا
کہ	کی	رہے	ہو
ہیں	کریں	اور	لگے
رکھتی	کیلئے	رکھتے	چلا

The produce for removing the stop words is summarized depicted in Table 3.6.

[1]https://www.ranks.nl/stopwords/urdu

Table 3.6 Algorithm for Removing Stop words

Input: Tokenize Text **Output:** Clean text with-out stop words
Using the list to store the result of output For loop • Comparing each word of the text with the words in the list of stop words • If matches with the list of stop words • Deleting the words • else • Storing it in the txt_clean Returning **Clean Text**

An example is given to illustrate the stop words removal process inTable3.7. It is clearly shown that after applying the algorithm of stop words on the news tweet words is سے, کرے,گ removed from the text.

Table 3.7 An example after stop words removal

	Text Data
Processed News	عالمی بینک عسکریت پسندی سے متاثرہ خاندانوں کی معاونت کرے گا
Stop words	'عالمی', 'بینک', 'عسکریت', 'پسندی', 'متاثرہ', 'خاندان', 'معاونت', 'گانا

3.2.3 Lemmatization

A corpus of words is reduced down to its lemma via the process of lemmatization (root word). The "lemma" of a word is determined by the process of lemmatization, which involves first determining the word's Parts of Speech and then examining its context within a particular phrase. Normalization is achieved by the process of lemmatization, which is used in the process of data retrieval applications

such as searching and indexing. In their paper, (Dave, Riddhi and Prem, 2015)examined a total of five different lemmatization strategies, including the Edit Distance on Dictionary algorithm, the Morphological analyzer, the radix tree, the Affix lemmatizer, and the fixed length truncation approach. The authors make an effort to reduce the significance of this impact by using normalization. We used Urdu Hack[2] library for this task. The algorithm which is used for the lemmatization of words is show in Table 3.8.

Table 3.8 Procedure for Lemmatization

Input: tokenizetext, lookup path **Output:** returning root words of a sentence
All the root words are defined in the Urdu hack library Using the lookup function of Urdu hack library to get the root words. **If words match with the word2lemma** **Return root word** Else **Return actual word**

The result of the lemmatization is provided in Table 3.9.

Table 3.9 Lemmatization Example

	Data
Input News	کے الیکٹرک کو اضافی بجلی گیس کی فراہمی کے قانونی تقاضے تعطل کا شکار
Word Lemmatization	الیکٹرک', 'کو', 'اضافی', 'بجلی', 'گیس', 'فراہمی', 'قانونی', 'تقاضا', ''تعطل', 'کا', 'شکار

The presented example demonstrates the application of lemmatization algorithm. The word تقاضے is converted to its root word which is تقاضا.

3.3 LIBRARIES

[2] https://urduhack.readthedocs.io/en/stable/reference/

3.3.1 Natural Language Processing (NLP)

The fields of Artificial Intelligence (AI), Computer Science (CS), and Industrial Engineering (IE) each provide a subfield of study that focuses on Natural Language Processing (NLP). It's a means via which humans and computers may exchange information and ideas. It is a piece of software that is capable of managing and analyzing vast volumes of data pertaining to natural language (Wikipedia Contributor, 2022). According to Seif and George (2017) this is the most effective method for bringing computers and people together at the moment since computers are unable to understand the feelings that humans experience. It was self-evident that NLP required development given that computers can multitask more effectively than people. On the other hand, recent advancements in the field of machine learning (ML) have made it possible for computers to carry out a broad variety of activities using natural language. We are now able to develop applications that are able to translate across languages, grasp the meaning of information, and summarizing it for the reader all via the use of machinelearning. These characteristics are useful in the sense that they remove the need for us to manually read and carry out computations on enormous amounts of text.

3.3.2 Natural Toolkit Library (NLTK)

Bird, Edward, and Loper(2009)The NLTK acronym stands for "natural language toolkit." English-language applications written in Python have the ability to make use of a wide number of libraries, including those dedicated to symbolic and statistical natural language processing. A few of the categories that are included in this toolkit are as follows: sentiment analysis, metrics, parsing, tagging, tokenization, chatting, chunking, categorizing, translating, Twittering, an interface, sketching, clustering, and a great deal of other things.

NLTK provides users with both visual representations of data as well as sample data. The language processing tasks that can be done by the toolkit are described in detail in a cookbook that is included in the package, as well as in a book that explains the principles of the language processing activities. Python was used in the development of the NLTK framework, which is a tool for doing statistical processing of natural languages on text written in natural languages (human language).

Due to the fact that it is an open-source project, the Natural Language Toolkit (NLTK)

is a Python library that is compatible with all operating systems. Inthesis, we used the NLTK toolkit to strip Urdu text of unnecessary stuff like URLs and special characters.

3.4 FEATURE EXTRACTION

In this section we are applying feature extraction algorithm on raw data. The process of transforming raw data into numerical features that may be processed while keeping the information in the original data set is referred to as feature extraction. It produces better outcomes than simply applying machine learning to raw data.

3.4.1 Bag of Words

The content of the text is converted into the count matrix by Bag of Words via the processes of tokenization and counting the occurrences of the feature (Garreta and Moncecchi, 2013). The phrases are tokenized, and then the characters of lengths higher than two are extracted from them. The existence of each feature is counted, and a sparse matrix is produced as a result.

Modeling text may be difficult because of the complexity of the data, and approaches like machine learning algorithms deal with inputs and outputs that are clearly defined and of a set length.

Text in its natural form cannot be fed directly into machine learning algorithms; instead, the text must be transformed into numbers. In particular, vectors of numerical values. The method is quite straightforward and adaptable, and it may be used in a wide variety of settings for the purpose of extracting characteristics from a document.

A depiction of text called a bag-of-words depicts the occurrence of words inside a document. It consists of two properties:

- A collection of words that are already known.
- A metric that indicates how many familiar terms are present.

This kind of document is referred to as a bag of words since it does not preserve any information on the sequence or arrangement of the words contained inside it. The model is primarily concerned with whether or not known terms appear in the content; it does not take into account where in the document the words appear. Some of the occurrence of Urdu word shown in Table3.10.

Table 3.10 Frequency of Some Words Used in Research

Words	Frequency	Words	Frequency
قومی	2080	کو	543
رکنا	71	میچ	4598
سیریز	89	مائیک	1979
خلاف	3164	اندراج	3473
میں	4803	مقدمہ	3827

3.4.2 Term Frequency-Inverse Document Frequency

TD-IDF is a statistical assessment of a term's significance to a corpus document. The term's significance increases with its occurrence in the text, but decreases as its frequency in the corpus rises. It has two terms: term frequency (TF) and inverse document frequency (IDF).

$$TF \times IDF \tag{1}$$

The frequency with which a phrase appears in a given text is measured using TF. The frequency is then divided by the whole length of the text so that the results may be normalized. Because bigger documents have a larger frequency of a phrase, normalization helps minimize bias toward bigger documents, which would otherwise be more favorable. The word frequency method assigns the same level of significance to each and every phrase in the text.

$$TF = \frac{Frequency\ of\ term\ t}{length\ of\ the\ document} \tag{2}$$

Most of the time, the terms that are not very important have greater frequency. So, words that are used a lot have to be weighed down. While the less-used but more important words need to be scaled up. Using IDF, we are scaling up the less frequentterms(Garreta et al.,2013).

$$IDF_t = \log\left(\frac{number\ of\ documents}{number\ of\ documents\ with\ term\ t}\right) \tag{3}$$

3.5 DIFFERENT APPROACHES USED IN LITERATURE

According to Hardeniya et al. (2016), approaches for textcategorizationcan be classified into three primary categories.

1. Machine Learning Approaches
2. Lexicon-based Approaches
3. Deep Learning Approaches

3.5.1 Machine Learning Approaches

According to Khan et al., (2022), To identify attitudes, this system of machine learning employs machine learning techniques and artificial intelligence. The forecasts are generated using a trained data set. The machine learning algorithm transforms text input into vectors based on the polarity of the phrase and looks for a preset pattern associated with each vector that is neutral, positive, or negative. Through the input of data, the system develops understanding and begins to anticipate classifications. By supplying increasingly precise data sets, the system's accuracy increases. The two forms of machine learning are further split into supervised and unsupervised learning.

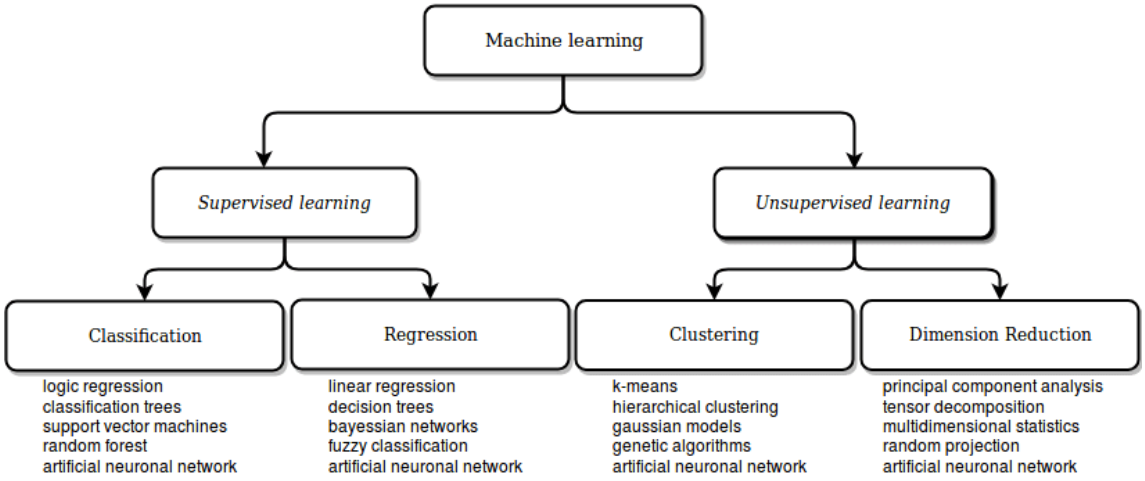

Figure 3.2 Approaches of Machine Learning

Machine learning approaches. (2018). Research Gate. https://www.researchgate.net/figure/Machine-learning-approaches-cited-in-38_fig2_324073224

3.5.1.1 Supervised Learning

Supervised learning is a specific kind of machine learning and artificial intelligence. It is characterized by the training of classification or prediction algorithms using labelled datasets. During the cross-validation phase, the model makes fine-tuned adjustments to its weights when new data is introduced into it. Organizations may use supervised learning to find large-scale solutions to a wide range of real-world challenges, including spam classification and removal from inboxes(IBM,n.d.)

In supervised learning, models are taught to produce the intended result by being exposed to a training set. This dataset is used to teach the model, since it contains both incorrect and right inputs and outputs. The algorithm monitors its performance using the loss function and makes adjustments until the error is suitably reduced.

In data mining, classification and regression are the two main categories of supervised learning issues.

- **Classification** employs an algorithm to precisely classify test results. It finds instances of known items within the dataset and makes educated guesses about how those things need to be categorized or defined. Some of the most popular classification techniques, including linear classifiers, support vector machines (SVM), decision trees, k-nearest neighbor, and random forest.
- **Regression** is a method for analyzing the connection between two variables. It is often used for forecasting purposes, especially with regards to the anticipated volume of sales for a particular enterprise. Well-known regression techniques include the linear, logistic, and polynomial varieties.

3.5.1.2 Unsupervised Learning

Unsupervised learning is a kind of machine learning that does not rely on labelled data. This information is then used to find patterns and associations that may aid in the resolution of clustering and association difficulties. This comes to be handy when domain specialists cannot identify general tendencies within a dataset collection. The hierarchical, k-means, and Gaussian mixed models clustering methods are often used.

The unsupervised learning algorithm can be further categorized into two types of problems:

- **Clustering**: Clustering is a strategy for organizing things into groups called clusters in such a way that the objects that share the most similarities stay together in one group while the objects that share the less or the no similarities go into another group. Cluster analysis is used to determine the similarities that exist between different data items and then classifies those data objects according to whether or not those similarities are present.
- **Association**: An unsupervised learning strategy that is used for the purpose of determining the connections that exist between the many variables included in a huge database is referred to as an association rule. It determines the group of objects that appear together in the dataset in a certain way.

3.5.2 Lexicon-Based Approaches

According to Jurek et al., (2015) One of the primary methods for performing sentiment analysis is the use of a lexicon, which entails assessing the sentiment based on the semantic orientation of words or phrases that appear in a text. With this strategy, a lexicon of words with positive and negative connotationseach with a corresponding positive or negative emotion valueis needed. In lexicon-based techniques, a text message is often represented as a collection of words. All positive and negative words or phrases are given emotion values from the lexicon after this depiction of the message. To combine data, a combining function, such as sum or average, is used. Machine learning models' primary drawback is their dependency on labelled data. Making sure that enough data can be collected and that it is appropriately labelled is quite tough. In addition, it is thought that a lexicon-based method is preferable for task since it is simpler for a person to comprehend and alter.

Bonta et al., (2019) use their context-free semantic orientation and, several lexicons, such as GI (General Inquirer) and LIWC (Linguistic Inquiry and Word Count), to classify words as positive or negative. Nearly 4,500 words make up LIWC, which are divided into 76 categories. Of them, 905 words specifically pertain to sentiment analysis in two categories. Despite being widely used to detect sentiment analysis in social media text, LIWC does not take into account emoticons, acronyms, initialisms, or slang, which are crucial elements for sentiment analysis of social media text.

3.5.3 Deep Learning Approaches

Heaton et al., (2017) declared that deep learning has been used to build investment portfolios, choose, and rate assets, and implement active risk management with success. Souma et al., (2019) states that high-frequency stock market prediction is a promising use of deep learning because of the method's capacity to extract features from a vast collection of raw data without requiring previous knowledge of predictors. The algorithms' success is highly dependent on the manner of data representation, and these choices of network topology, activation function, and other model parameters vary widely. The benefits and pitfalls of deep learning algorithms have been studied by researchers.

Naqvi et al., (2021) stated that for difficult issues like speech recognition, picture classification, and natural language processing, deep learning offers outstanding advantages. Deep learning is a preferable alternative over traditional machine learning techniques for text classification. Recurrent neural networks and convolutional neural networks, two deep learning techniques are outperformed classifiers that relied on human feature creation. As Deep Learning methods use word embedding as inputs, which already lists the semantic data and learns the relation between data elements in the sentence. So, there is no need to include rules or special features along with the text.

Deep learning techniques, used by Akhter et al., (2020) for the categorization of Urdu documents in the production of goods. In doing so, they lowered accuracy for big datasets and boosted accuracy for small and medium-sized datasets by eliminating stop words and unusual terms.

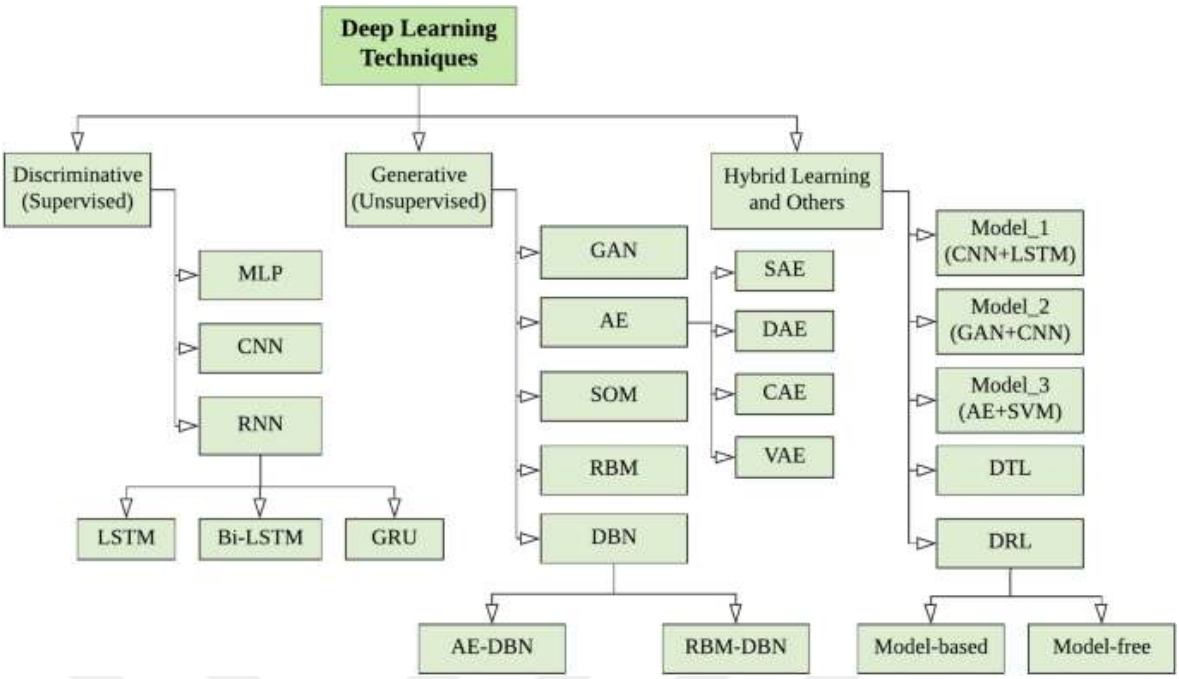

Figure 3.3 Approaches of Deep Learning

Deep Learning. (2021, August 18). Springer Link. https://link.springer.com/article/10.1007/s42979-021-00815-1/figures/6

3.6 APPROACHES USED IN EXPERIMENTS

In this study, we employ traditional machine learning approachesto categorize the Urdu news text. Algorithm we used for the categorization problem are as follows:

1. Logistic Regression
2. Multinomial Naïve Bayes
3. Support Vector Machine
4. Decision Tree Classifier

3.6.1 Logistic Regression

Logistic regression is a kind of learning technique for classification that is used under supervision to make predictions about the likelihood of a target variable. The nature of the target, also known as the dependent variable, is dichotomous, which

indicates that there are only two conceivable categories(*Machine Learning - Logistic Regression*, n.d.).

The dependent variable is of a binary type, meaning that its data may be encoded as either a 1 (which represents for success/yes) or a 0 (which stands for failure/no).

A logistic regression model makes a mathematical prediction of P(Y=1) as a function of the variable X. It is one of the simplest machine learning methods, and it may be used to a wide variety of classification issues, including the detection of spam, the prediction of diabetes, the detection of cancer, and the categorization of news.

In most cases, the term "logistic regression" refers to a binary logistic regression in which the target variables are also binary. However, logistic regression may also predict the values of two additional categories of target variables. Logistic regression may be broken down into the following categories and subtypes based on the number of categories involved:

3.6.1.1Binomial

In this particular method of categorization, a dependent variable will only be able to take on one of two potential values, either 1 or 0. For instance, these variables may indicate a successful a yes or a no, a win or a loss, and so on.

3.6.1.2Multinomial

In this form of categorization, the dependent variable may include three or more alternative categories that are not ordered, as well as types that have no quantitative importance. These variables, for instance, may be used to represent "Category A," "Category B," or "Category C."

3.6.1.3Ordinal

In this form of categorization, the dependent variable might have three or more potential ordered categories, or the types can have a quantitative importance. For instance, these variables may indicate "bad," "good," "very good," or "excellent," and each category might contain scores such as 0, 1, 2, or 3.

One of the most popular and effective methods for solving classification issues is logistic regression. Logistic regression relies on hypothesis functions that consistently provide predictions in the range 0–1.

$$\theta \leq h_\theta(\theta^T x) \leq 1 \tag{4}$$

A sigmoid function is used to represent the hypothesis function.

$$h_\theta(\theta^{\mathrm{T}} x) = \frac{1}{1 + e - \theta^{\mathrm{T}} x} \tag{5}$$

For logistic regression, x is the input variable or feature, and $h_\theta(\theta^{\mathrm{T}} x)$ is the hypothesis function with a parameter ofθ.

The sigmoid curve can be illustrated with the help of Figure 3.4. The values of y-axis lie between 0 and 1 and crosses the x-axis at 0.5.

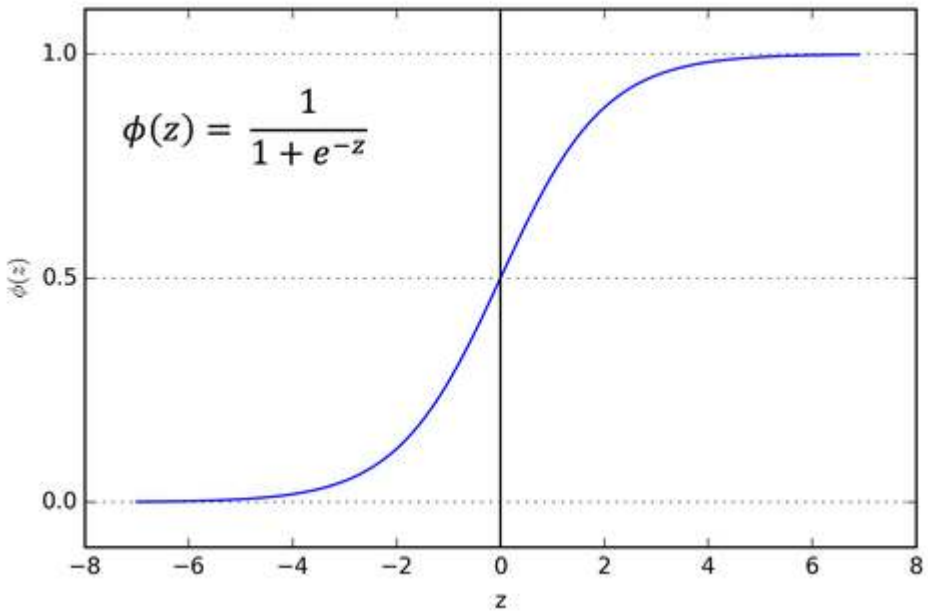

Figure 3.4 Sigmoid Curve

Jang, J. (2021, September 16). *Sigmoid Function*. medium.com. https://medium.com/@shiny_jay/logistic-regression-a9a8749e1e68

The classes can be labelled as positive or negative. If the result is between 0 and 1 the outcome belongs to positive class with high probability.According to our assumption, we interpret the output of hypothesis function as positive if it is ≥0.5,

otherwise negative.

In this work, we used the logistic regression with Bag of words (BOW) and the Term Frequency Inverse Document Frequency(TF-IDF). First, we preprocess the data and then afterthat we split the data between training and testing. For training we are using 80% and testing 20% and then fed it to the model. After that we are using the predict function to predict the accuracy and we are getting the accuracy of 0.88250 through logistic regression with bow and 0.88875 through TF-IDF.

3.6.2 Naïve Bayes

Lagerkrants and Holmström (2016),The Naïve Bayes model belongs a family of classifiers that makes what is referred as naive assumption. The naive assumption is that attributes(words) are independent of each other.

Scikit learn which is one of the most useful libraries of Python,will be used to construct a Naïve Bayes model. There are three types of Naïve Bayes model under Scikit learn:

3.6.2.1 Gaussian Naïve Bayes

The assumption that the data from each label is obtained from a simple Gaussian distribution makes this the most basic form of the Naive Bayes classifier(*Classification Algorithms - Naive Bayes*, n.d.)

3.6.2.2 Multi-Nomial Naïve Bayes

The Multinomial Naive Bayes classifier is yet another helpful use of the Naive Bayes algorithm. In this variant, the features are said to be derived from a straightforward multinomial distribution. This particular kind of Naive Bayes is best effective when used to features that reflect discrete counts(*Classification Algorithms - Naive Bayes*, n.d.)

3.6.2.3Bernoulli Naïve Bayes

Bernoulli Naive Bayes is another significant model. In this model,

characteristics are believed to be either binary or continuous (0s and 1s). One possible use of the Bernoulli Naive Bayes algorithm is text categorization using the "bag of words" paradigm(*Classification Algorithms - Naive Bayes*, n.d.)

The Naïve Bayes model we used is in our experiments is commonly called a Multinomial Naïve Bayes model. This model takes into account the number of occurrences of an attribute in a document. The Naïve Bayes framework is provided by a simple theorem of probability known as Baye's rule:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \tag{6}$$

3.6.3 Support Vector Machines

It has been shown that support vector machines are quite good in conventional text classification. The Support Vector Machine (SVM) is a technique for supervised machine learning that can categories data and may also be used to solve regression issues. The majority of the time, it is used in the challenges involving categorization. When working with SVM, each feature of our dataset is represented as a point in our coordinate system. The method looks for a hyper-plane that can divide the two classes with the highest degree of precision feasible and attempts to identify it.

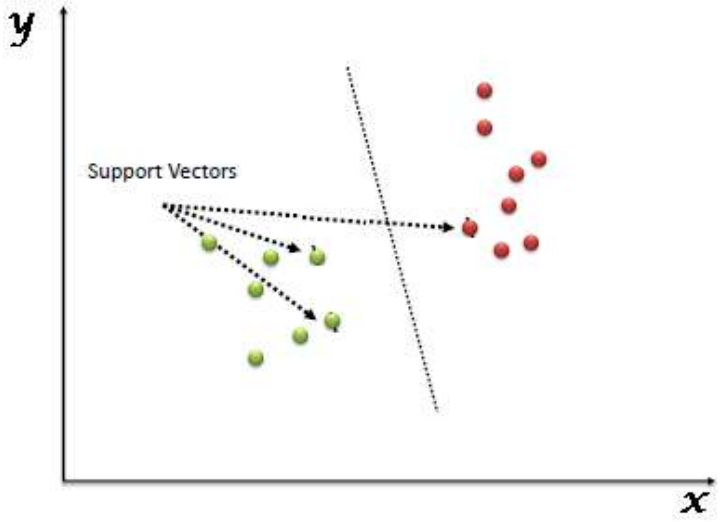

Figure 3.5 Hyper plane Separating Two Classes

Ray, S. (2017, September 13). *SVM Hyperplane*. analyticsvidhya.com.

https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

The term "margin" refers to the space that separates the points that are considered to be the closest in each of the categories. It is necessary to choose a hyperplane that is capable of determining the highest possible margin for both of the categories. If the data are nonlinear, a specialized method known as kernelization is used to analyze them. During this stage of the procedure, the data are first projected onto a three-dimensional plane, and after that, a hyperplane that divides the two groups is identified. After that, the points are re-projected onto the two-dimensional space. Therefore, it is possible that the hyperplane has a circular shape (Sreedevi and Rama Bai, 2020).

We used Support Vector Machines (SVM), which is another well-known classifier. Firstly, an instance of the SVC Classifier has been generated. Then, we repeat the process of training the model with training data and predicting classes for the test data. Prior to the construction of SVM model for the categorization of news, the first step is to eliminate stop words, and after that, punctuation marks are eliminated. In the next step, is to clean the numbers, URLs, and other special characters since these things do not contribute to the classification of the text. After all of the data has been preprocessed, the next step is to utilize bag of words to produce sparse matrix. This sparse matrix will include the words as vectors, and they will be counted as the feature. The support vector machine classifier is used here to analyze this data. The finding's part includes an analysis and discussion of the results, as well as the implications that can be drawn from them.

3.6.4 Decision Tree Classifier

Applying the techniques of decision tree analysis, a kind of predictive modelling, may help in a wide variety of settings. Using an algorithmic method, the dataset may be partitioned in a variety of ways depending on the criteria being used to form the decision tree. In the class of algorithms known as "supervised algorithms," decision trees stand out as the most effective. They are versatile and may be used for both classification and regression. The decision nodes at which the data is partitioned and the resultant leaves are the two primary components of a tree(*Classification Algorithms - Decision Tree*, n.d.)

The information gains of each of the characteristics are calculated in a decision tree, and the feature that results in the greatest gain is selected to serve as the tree's root node. In a similar manner, the information gains for the remaining features are computed, the features for the next level of the tree are determined, and the tree is constructed by repeatedly following the same approach

4 FINDINGS

This section contains the results of the machine learning algorithm which we used to perform the Urdu news categorization and the comparison of all the models.

4.1 PERFORMANCE MEASURES

The data was split into two parts with 80 percentages for training and 20 percentages for testing purpose. 20% data was used as unseen data that was used to test the performance of different classifiers as depicted in Table (4.1). We evaluate the effectiveness of our models using Recall (R), Precision (P), and F1-measure. The mathematical equations are as follows:

Table 4.1 Data Splitting

Data Split	Number of News	Percentage
Train	3200	80%
Test	800	20%
Total	4000	100%

4.1.1 Accuracy

The number of right predictions that a model generates in comparison to the

total number of predictions generated by the model is accuracy. The more precise the model, the better it will perform.

$$Accuracy = \frac{\text{Number of true predictions}}{\text{Size of the test dataset}} \quad (7)$$

4.1.2 Precision

Precision is the proportion of positive instances accurately identified by the model relative to the total number of positive instances identified by the model. The formula to calculate the precision.

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

Where TP and FP stand for true positive and false positive.

4.1.3 Recall

Recall measures how many positive cases were properly identified relative to the total number of positive instances in the test dataset.

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

Where TP and FN stand for true positive and false negative.

4.1.4 F-Score

The F-score is calculated by taking the geometric mean of the model's accuracy and recall. When the value of the F-Score is higher, the categorization results become more accurate.

$$F_1 = \frac{2 * P * R}{P + R} \quad (10)$$

Where P and R stand for true Precision and Recall which we discussed above.

4.2 PERFORMANCE OF NAÏVE BAYES

The Naïve Bayes model was implemented in Python language. The result of the experiment was noted in terms of confusion matrix (Table 4.2,4.3), which was used to calculate three (accuracy, precision, and recall) performance metrics (Table4.4).

Table 4.2 Confusion Matrix of Training Set of Naïve Bayes

	Business&Economics	**Entertainment**	**Science&Technology**	**Sports**	
Business&Economics	**781**	3	23	3	810
Entertainment	9	**777**	15	11	812
Science&Technology	17	4	**774**	2	797
Sports	5	10	3	**763**	781
Total	812	794	815	779	**3200**

Table 4.3 Confusion Matrix of Testing Set of Naïve Bayes

	Business&Economics	**Entertainment**	**Science&Technology**	**Sports**	
Business&Economics	**176**	2	8	4	190
Entertainment	8	**158**	11	11	188
Science&Technology	7	7	**185**	4	203
Sports	9	8	2	**200**	219
Total	200	175	206	219	**800**

Here we have a 4x4 confusion matrix. The number of occurrences that have been successfully identified is equal to the sum of the diagonal elements in the confusion matrix; all other examples have been incorrectly categorized. For the sake of

calculation, let TP_B represent the number of true positives in the Business&Economics class, TP_E represent the number of true positives in the Entertainment class, and TP_{ST} represent the number of true positives in the Science and Technology class. TP_S is the number of class Sports true positives.

- TP_B: describes the 176 matches in the first row and column that the classifier got right (Business and Economics).

- TP_E: refers to the 158 affirmative tuples in the second row, second column that were properly labelled as Entertainment by the classifier.

- TP_{ST}: refers to the 185 matches in the third row and third column that the classifier properly categorized as belonging to the Science and Technology category.

- TP_S: defines the two hundred tuples in the fourth row, fourth column that the classifier got right and labelled (Sports).

$$Accuracy = \frac{TP_B + TP_E + TP_{ST} + TP_S}{Total\ instance\ of\ class} \tag{11}$$

$$Accuracy = \frac{176+158+185+200}{800} = 0.90$$

In the same way, the total number of wrongly classified cases are the cases in the confusion matrix table that are not highlighted. When you divide the total number of that cases by the total number of classified cases, you get the number of wrongly classified cases.

The most confusing categories using Naïve Bayes are Business&Economics and Science&Technology.

Table 4.4 Performance of Naïve Bayes

	Precision	Recall	F1-score
Business&Economics	0.88	0.93	0.90
Entertainment	0.90	0.84	0.87
Science & Technology	0.90	0.91	0.90
Sports	0.91	0.91	0.91
Accuracy			0.90

4.3 PERFORMANCE OF LOGISTIC REGRESSION

The performance of the Logistic regression model was assessed by this experiment. The trial results are provided in terms of the confusion matrix (Tables 4.5, 4.6); these findings are used to calculate accuracy, precision, and recall (Table 4.7).

Table 4.5 Confusion Matrix of Training Set of Logistic Regression

	Business&Economics	Entertainment	Science&Technology	Sports	
Business&Economics	**741**	69	0	0	810
Entertainment	0	**812**	0	0	812
Science&Technology	10	330	**457**	0	797
Sports	1	334	0	**446**	781
Total	752	1545	457	446	**3200**

Table 4.6 Confusion Matrix of Testing set Logistic regression

	Business&Economics	Entertainment	Science&Technology	Sports	
Business&Economics	**144**	45	0	1	190

Entertainment	0	**188**	0	0	188
Science&Technology	3	101	**99**	0	203
Sports	5	115	0	**99**	219
Total	152	449	99	100	**800**

We have a 4x4 confusion matrix. The number of occurrences that have been successfully identified is equal to the sum of the diagonal elements in the confusion matrix; all other examples have been incorrectly categorized. For the sake of calculation, let TP_B represent the number of true positives in the Business&Economics class, TP_E represent the number of true positives in the Entertainment class, and TP_{ST} represent the number of true positives in the Science and Technology class. TP_S is the number of true positives instances of class Sports.

- TP_B: corresponds to the 144 positive tuples successfully identified by the classifier in the first row-first column (Business and Economics).

- TP_E: corresponds to the 188 positive tuples successfully classed as Entertainment by the classifier in the second row - second column.

- TP_{ST}: corresponds to the 99 positive tuples successfully identified (Science & Technology) by the classifier in the third row-third column.

- TP_S: corresponds to the 99 positive tuples successfully labelled (Sports) by the classifier in the fourth row-fourth column.

Accuracy can be computed by the below equation:

$$Accuracy = \frac{TP_B + TP_E + TP_{ST} + TP_S}{Total\ instance\ of\ class}$$

$$Accuracy = \frac{144+188+99+99}{800} = 0.88$$

In Table 4.7, the incorrectly identified cases are not highlighted. Total instances divided by total categorized instances gives incorrect instances.

Logistic Regression performs worse predicting the instances in the entertainment subcategory due to the fact that the news about the entertainment industry is also somehow belongs to both science and technology news and sports news.

Table 4.7 Performance of Logistic Regression

	Precision	**Recall**	**F1-score**
Business&Economics	0.87	0.91	0.89
Entertainment	0.83	0.88	0.85
Science & Technology	0.93	0.90	0.91
Sports	0.89	0.85	0.87
Accuracy			0.88

4.4 PERFORMANCE OF SVM

Python was used for the implementation of the SVM model. The outcome of the experiment was recorded using the confusion matrix (Tables 4.8 and 4.9), which was used to construct three performance measures (accuracy, precision, and recall) (Table 4.10).

Table 4.8 Confusion Matrix of Training Set of SVM

	Business&Economics	Entertainment	Science&Technology	Sports	
Business&Economics	**803**	0	7	0	810
Entertainment	0	**809**	2	1	812
Science&Technology	0	1	**796**	0	797
Sports	0	6	0	**775**	781
Total	803	816	805	776	**3200**

Table 4.9 Confusion Matrix of Testing set SVM

	Business&Economics	Entertainment	Science&Technology	Sports	
Business&Economics	**169**	9	6	6	190
Entertainment	9	**172**	5	2	188
Science&Technology	6	19	**175**	3	203
Sports	7	32	2	**178**	219
Total	191	232	188	189	**800**

Here is a 4x4 confusion matrix. The sum of the diagonal elements in the confusion matrix is the number of cases that were correctly classified. All the other cases were wrongly classified. Let's say that TP_B is the number of true positives in the Business and Economics class, TP_E is the number of true positives in the Entertainment class, and TP_{ST} is the number of true positives in the Science and Technology class. TP_Sdenotes the true positive of class Sports.

- TP_B: indicates to the positive tuples in the first row and first column, that the classifier has properly identified as (Business and Economics) are 169.

- TP_E: indicates to the positive tuples in the second row and second column

column, that the classifier has properly identified as (Entertainment) are 172.

- TP_{ST}: indicates to the positive tuples in the third row and third column, that the classifier has properly identified as (Science and Technology) are 175.

- TP_S: indicates to the positive tuples in the fourth row and fourth column, that the classifier has properly identified as (Sports) are 178.

As a result, the equation may be used to determine the accuracy of the cases that were properly identified.

$$Accuracy = \frac{TP_B + TP_E + TP_{ST} + TP_S}{Total\ instance\ of\ class}$$

$$Accuracy = \frac{169+172+175+178}{800} = 0.86750$$

Parallel to this, the occurrences other than the ones marked in the confusion matrix table make up the total wrongly classified instances. You may determine the number of incorrectly classified cases by dividing the whole sum of those occurrences by the total number of classified instances.

Using SVM when we tested the predicted results it get confused predicting the Entertainment news with Science&Technology and Sports news.

Table 4.10 Performance of SVM

	Precision	**Recall**	**F1-score**
Business&Economics	0.88	0.89	0.89
Entertainment	0.74	0.91	0.82
Science & Technology	0.93	0.86	0.90
Sports	0.94	0.81	0.87
Accuracy			0.87

4.5 PERFORMANCE OF DECISION TREE

Python was used to build the Decision Tree model. The result of the experiment was written down in the confusion matrix (Table 4.11 and Table 4.12), which was used to figure out three performance metrics: accuracy, precision, and recall (Table 4.13).

Table 4.11 Confusion Matrix of Training Set of Decision Tree

	Business&Economics	**Entertainment**	**Science&Technology**	**Sports**	
Business and Economics	**810**	0	0	0	810
Entertainment	0	**812**	0	0	812
Science and Technology	0	0	**797**	0	797
Sports	0	0	0	**781**	781
Total	810	812	797	781	**3200**

Table 4.12 Confusion Matrix of Testing set Decision Tree

	Business&Economics	**Entertainment**	**Science&Technology**	**Sports**	
Business and Economics	**152**	19	14	5	190
Entertainment	13	**157**	8	10	188
Science and Technology	15	15	**170**	3	203
Sports	12	31	4	**172**	219
Total	192	222	196	190	**800**

Here we have 4x4 confusion matrix. If you add up the elements on the diagonal of the confusion matrix, you get the total number of correctly classified occurrences. As for the purpose of calculations, Let TP_Bbe the value of true positives in Business and Economics, TP_E in Entertainment, and TP_{ST} in Science and Technology and TP_Sin class

Sports'.

- TP_B: denotes to the 152 positive tuples properly identified by the classifier in the first row-first column(Business&Economics).

- TP_E: denotes to the 157 positive tuples properly identified by the classifier in the second row – second column (Entertainment).

- TP_{ST}: denotes the 170 positive tuples successfully identified (Science & Technology) by the classifier in the third row-third column.

- TP_S: denotes the 172 positive tuples successfully identified (Sports) by the classifier in the fourth row-fourth column.

As a result, the accuracy of the instances that have been properly categorized is computed by using the below equation.

$$Accuracy = \frac{TP_B + TP_E + TP_{ST} + TP_S}{Total\ instance\ of\ class}$$

$$Accuracy = \frac{152+157+170+172}{800} = 0.81$$

In the confusion matrix table, the incorrectly identified cases are those which are not highlighted.The number of incorrectly categorized instancescalculated by taking the whole sum of such cases and dividing it by the total number of classified instances.
While using the decision tree, the Entertainment category was the most difficult to understand.

Table 4.13 Performance of Decision Tree

	Precision	**Recall**	**F1-score**
Business&Economics	0.79	0.80	0.80
Entertainment	0.71	0.84	0.77
Science & Technology	0.87	0.84	0.85

Sports	0.91	0.79	0.84
Accuracy			0.81

4.6 WEB-APPLICATION INTERFACE

Web interface is developed in Python using flask. First we saved the model using pickle library and then we build the HTML page. Through HTML page we are getting the news headline and the news body as an input from users and then model returns it predicted category. Figure 4.5 depicts the web interface developed.

4.7COMPARISON OF ALL THE MODELS

For this study we use 4 different models' Logistic regression, naïve bayes, SVM, and decision tree and for each model we are computing the accuracy, F1-support, precision and the recall. Through naïve bayes we are achieving the precision 0.898539, recall 0.897828 and F1-score 0.897744. With SVM we are getting the precision 0.874712, recall 0.869805 and F1-score 0.868469 and with Logistic model we are getting the precision 0.839019, recall 0.674409 and F1-support 0.677173 and with decision tree we are obtaining the result of precision 0.817871, recall 0.814483 and F1-support 0.813718. All of these results are justified with the graphs of models mentioned in Figures (6,7,8) below and also with Table4.14.

Table 4.14 Comparison of the Classifiers

Models	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.88250	0.839019	0.674409	0.677173
SVM	0.86750	0.874712	0.869805	0.868469
Naïve Bayes	**0.89875**	**0.898539**	**0.869805**	**0.897744**
Decision Tree	0.79125	0.817871	0.814483	0.813718

For Naïve Bayes we are achieving the accuracy of **0.89875** and with the model SVM we are getting **0.86750** and with Logistic model we are achieving the accuracy of **0.88250** as show in the Figure 9 and Table 23. From all of these classifiers Naïve Bayes is outperforming well with the accuracy of **0.89875.**

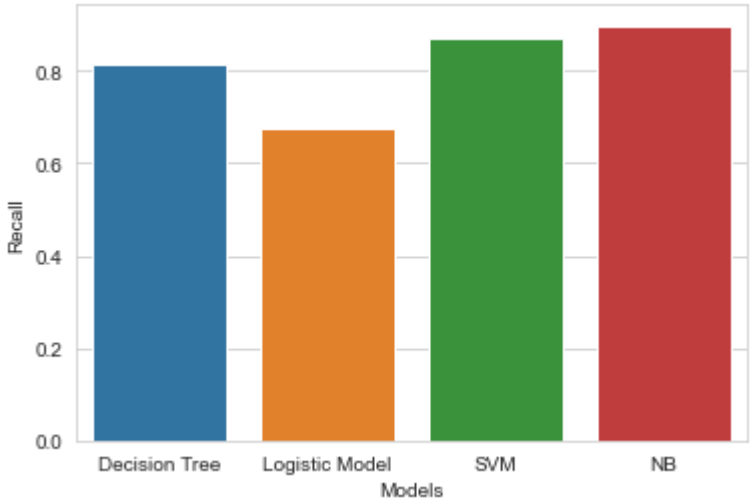

Figure 4.1 Recall Comparison

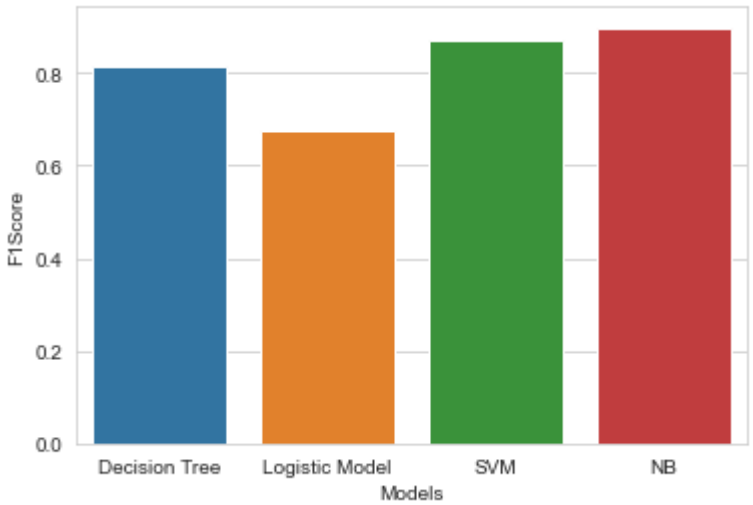

Figure 4.2 F1-Score Comparison

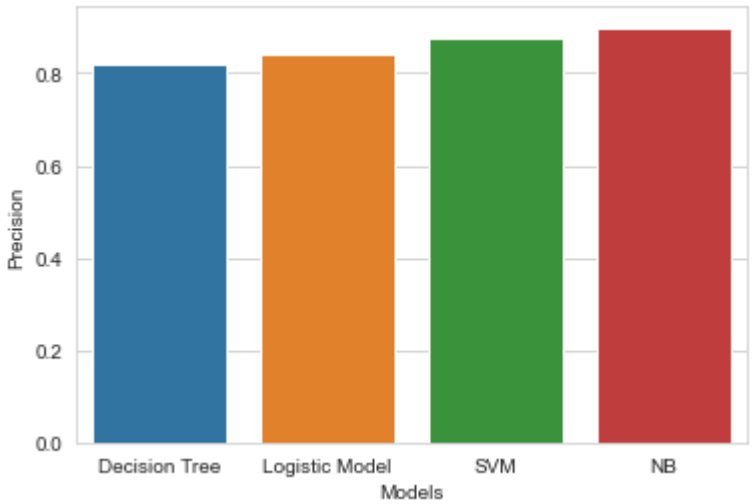

Figure 4.3 Precision Comparison

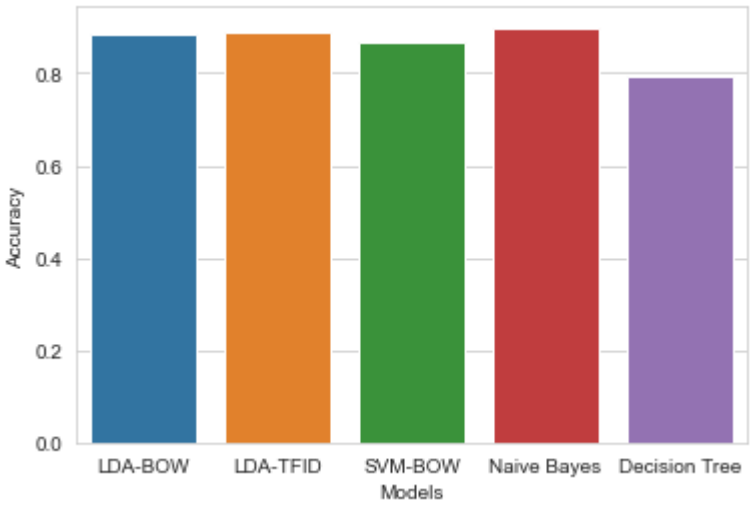

Figure 4.4 Accuracy Comparison of Model

Figure 4.5 Web-Application Interface

5CONCLUSION

This thesis presents a model for Urdu text classification. The data has been collected from an Urdu news website in different categories. We have employed Natural Language Toolkit and state of the art machine learning algorithms to classify each news text into predefined categories.

Furthermore, to get rid of the noise in the Urdu text before processing it or filtering it, a series of pre-processing processes must be performed. We remove the URL, links, and special characters from the text in these stages. To conduct fundamental natural language processing on the Urdu text, we will be using the NLTK toolkit function WordPunctTokenizer to perform the tokenization, which allows us to tokenize each word in the phrases. Following that, we utilized the Urdu hack library to accomplish word lemmatization. After preprocessing, the data is cleaned.

The goal of this study was to evaluate the performance of conventional machine learning methods such as Naïve Bayes, SVM, Logistic Regression and Decision Tree. For this data set and text classification problem, the Naïve Bayes classifier performs the best whereas Decision Tree performs the worst.

5.1 FUTURE DIRECTIONS

The findings of this thesis could be extended in several directions. The list below suggests some open problems for further study.

- The sentiment analysis problem can also be addressed with the text categorizations. This requires proposing a two-stage classifier.
- Deep learning models can be utilized provided that more data is collected.
- Although this thesis studies deals with static data, a system can be developed which performs real-time categorization.

REFERENCES

Al-Anzi, F. S., &AbuZeina, D. (2018). Beyond vector space model for hierarchical Arabic text classification: A Markov chain approach. *Information Processing & Management*, *54*(1), 105-115.

Al-Harbi, S., Almuhareb, A., Al-Thubaity, A., Khorsheed, M. S., & Al-Rajeh, A. (2008). Automatic Arabic text classification.

Ali, A. R., &Ijaz, M. (2009, December). Urdu text classification. In *Proceedings of the 7th international conference on frontiers of information technology* (pp. 1-7).

Bhumika, P. S. S. S., &Nayyar, P. A. (2013). A review paper on algorithms used for text classification. *International Journal of Application or Innovation in Engineering & Management*, *3*(2), 90-99.

Bird, S., Klein, E., &Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Bonta, V., &Janardhan, N. K. N. (2019). A comprehensive study on lexicon-based approaches for sentiment analysis. Asian Journal of Computer Science and Technology, 8(S2), 1-6.,

Brownlee, J. (2019, August 7). *A gentle introduction to the bag-of-words model*. MachineLearningMastery.com.

Chy, A. N., Seddiqui, M. H., & Das, S. (2014, March). Bangla news classification using naive Bayes classifier. In *16th Int'l Conf. Computer and Information Technology* (pp. 366-371). IEEE.

Classification algorithms - logistic regression. Tutorials Point. (n.d.). Retrieved December 28, 2022, from https://www.tutorialspoint.com/machine_learning_with_python/classification_algorithms_logistic_regression.htm

Dalal, M. K., &Zaveri, M. A. (2011). Automatic text classification: a technical review. *International Journal of Computer Applications*, *28*(2), 37-40.

Duwairi, R., Al-Refai, M. N., &Khasawneh, N. (2009). Feature reduction techniques for Arabic text categorization. *Journal of the American society for information science and technology*, *60*(11), 2347-2352.

Garreta, R., &Moncecchi, G. (2013). *Learning scikit-learn: machine learning in python*. Packt Publishing Ltd.

Han, E. H. S., Karypis, G., & Kumar, V. (2001, April). Text categorization using weight adjusted k-nearest neighbor classification. In *Pacific-asia conference on knowledge discovery and data mining* (pp. 53-65). Springer, Berlin, Heidelberg.

Hardeniya, T., &Borikar, D. A. (2016). Dictionary based approach to sentiment analysis-a review. International Journal of Advanced Engineering, Management and Science, 2(5), 239438

Hassan, M. A. (2016). *A Framework to Improve Classification of Positive and Negative Opinions in Roman Urdu-English Code Switching Environment* (Doctoral dissertation, University of Engineering & Technology, Lahore.).

Heaton, J., Polson, N., & Witte, J. H. (2017). Deep learning for finance: deep portfolios. Applied Stochastic Models in Business and Industry, 33(1), 3–12

Hossain, M. R., Sarkar, S., & Rahman, M. (2020). Different Machine Learning based Approaches of Baseline and Deep Learning Models for Bengali News Categorization. *International Journal of Computer Applications*, *975*, 8887.

Irsan, I. C., &Khodra, M. L. (2016, August). Hierarchical multilabel classification for Indonesian news articles. In *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)* (pp. 1-6). IEEE.

Jurek, A., Mulvenna, M.D. & Bi, Y. Improved lexicon-based sentiment analysis for social media analytics. Secur Inform 4, 9 (2015)

Kamruzzaman, S. M., Haider, F., & Hasan, A. R. (2010). Text classification using data mining. *arXiv preprint arXiv:1009.4987*.

Khan, L., Amjad, A., Ashraf, N., & Chang, H. T. (2022). Multi-class sentiment analysis of urdu text using multilingual BERT. Scientific Reports, 12(1), 1-17

Khan, L., Amjad, A., Ashraf, N., & Chang, H. T. (2022). Multi-class sentiment analysis of urdu text using multilingual BERT. Scientific Reports, 12(1), 1-17

Krail, N., & Gupta, V. (2012, December). Domain based classification of Punjabi text documents using ontology and hybrid based approach. In *Proceedings of the*

3rd Workshop on South and Southeast Asian Natural Language Processing (pp. 109-122).

Lagerkrants, E., &Holmström, J. (2016). Using machine learning to classify news articles.

Li, B., Chen, N., Wen, J., Jin, X., & Shi, Y. (2015). Text categorization system for stock prediction. *International Journal of u-and e-Service, Science and Technology*, *8*(2), 35-44.

Logistic sigmoid. Logistic Sigmoid - an overview | ScienceDirect Topics. (n.d.). Retrieved December 28, 2022, from https://www.sciencedirect.com/topics/computer-science/logistic-sigmoid

M. P. Akhter, Z. Jiangbin, I. R. Naqvi, M. Abdelmajeed and M. Fayyaz, "Exploring deep learning approaches for Urdu text classification in product manufacturing", Enterprise Inf. Syst., pp. 1-26, May 2020

Machine Learning - Logistic Regression. (n.d.). https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_classification_algorithms_logistic_regression.htm

Menaka, S., &Radha, N. (2013). Text classification using keyword extraction technique. *International Journal of Advanced Research in Computer Science and Software Engineering*, *3*(12), 734-740.

Naqvi, A. Majid and S. A. Abbas, "UTSA: Urdu Text Sentiment Analysis Using Deep Learning Methods," in IEEE Access, vol. 9, pp. 114085-114094, 2021, doi: 10.1109/ACCESS.2021.3104308

Odeh, A., Abu-Errub, A., Shambour, Q., &Turab, N. (2015). Arabic text categorization algorithm using vector evaluation method. *arXiv preprint arXiv:1501.01318*.

Purohit, A., Atre, D., Jaswani, P., &Asawara, P. (2015). Text classification in data mining. *International Journal of Scientific and Research Publications*, *5*(6), 1-7.

Ramdass, D., &Seshasai, S. (2009). Document classification for newspaper articles. *Document classification for newspaper articles*.

Riddhi, D., Prem, B.: Survey paper of different lemmatization approaches. Int. J. Res. Advent Technol. (2015). (E-ISSN 2321-9637) Special Issue 1st International Conference on Advent Trends in Engineering, Science and Technology

Seif, G. (2018). An easy introduction to natural language processing. *Towards Data Science*.

Sreedevi, J., Rama Bai, M., & Reddy, C. (2020). Newspaper article classification using machine learning techniques. *Int. J. Innov. Technol. Explor. Eng*, 2278-3075.

Sreedevi, J., Rama Bai, M., & Reddy, C. (2020). Newspaper article classification using machine learning techniques. *Int. J. Innov. Technol. Explor. Eng*, 2278-3075.

Vijayarani, S., Janani, R.: Text mining: open source tokenization tools - an analysis. Adv. Comput. Intell. Int. J. (ACII) 3(1), 37–47 (2016)

What is Supervised Learning? | IBM. (n.d.). https://www.ibm.com/topics/supervised-learning

What is supervised learning? IBM. (n.d.). Retrieved December 26, 2022, from https://www.ibm.com/cloud/learn/supervised-learning

Wikipedia contributors. (2022, December 25). Natural language processing. In *Wikipedia, The Free Encyclopedia*. Retrieved 10:00, January 3, 2023, from https://en.wikipedia.org/w/index.php?title=Natural_language_processing&oldid=1129486584

Yogish, D., Manjunath, T. N., &Hegadi, R. S. (2018, December). Review on natural language processing trends and techniques using NLTK. In *International Conference on Recent Trends in Image Processing and Pattern Recognition* (pp. 589-606). Springer, Singapore.

BIOGRAPHY

Muhammad Talha Satti

Muhammad Talha Satti is a student of last year of Master's program (Computer Engineering) at Beykoz University. He received a Bachelor's degree in Software Engineering from The Capital University of Science & Technology. He is currently working as a freelancer on Upwork and providing services as a web-developer. He is interested in making videos of nature and travelling different countries.