|  |  |
| --- | --- |
| Project Report  **Urdu Headlines classification** | By  Rabel Safina Aatiqa Batool  Machine Learning |

Table of Contents

[**I.** **Abstract** 2](#_Toc96777968)

[**II.** **Introduction** 2](#_Toc96777969)

[**III.** **Methodology** 2](#_Toc96777970)

[**Flow Chart:** 3](#_Toc96777971)

[**a.** **Corpus Collection.** 4](#_Toc96777972)

[Fig : 1.1 4](#_Toc96777973)

[Fig: 1.2 4](#_Toc96777974)

[**b.** **Stop Words Removal.** 5](#_Toc96777975)

[Fig: 2.1 5](#_Toc96777976)

[**c.** **Feature Selection.** 5](#_Toc96777977)

[**IV.** **Results** 5](#_Toc96777978)

[Fig: 3.1 6](#_Toc96777979)

[**V.** **CONCLUSION** 6](#_Toc96777980)

[**VI.** **References** 6](#_Toc96777981)

1. **Abstract**

Nowadays on the Internet there are a lot of sources that generate immense amounts of daily Urdu news. In addition, the demand for information by users has been growing continuously, so it is crucial that the news is classified to allow users to access the information of interest quickly and effectively. This way, the machine learning model for automated news classification could be used to identify topics of untracked Urdu news and/or make individual suggestions based on the user’s prior interests. Thus, our aim is to build models that take as input Urdu news headline and short description and output Urdu news category.

Classifying text in Urdu is very challenging task, especially when we have multiple classes to implement in multiple machine learning algorithm.

Humans can recognize and correctly provide multiple relevant labels for an article, but can a machine learning system get similar results?

Our corpus contains 111860 new words of 4 categories with (Business & Economics, sports, Entertainment, Science & Technology)

Category classification, for news, is a multi-label text classification problem. The goal is to assign one or more categories to a news article. A standard technique in multi-label text classification is to use a set of binary classifiers.

1. **Introduction**

Nowadays data on Internet is available in all type of spoken languages which is easy to understand, and people can grab data accordingly. Urdu language is considered as national language in Pakistan and most spoken language when we focus on Indo-Pak regions. More than 100 million people speaks Urdu widely throughout world. There is a lot of work still required in the Urdu language specially in the area of classifying text into different categories which is one of the most common and useful technique use to solve problem like spam filtering which

Predict data either its spam or not. Another area in which this technique is mostly use is sentiment analysis where we can identify class especially negative and positive using dataset.

Text classification is helpful in article tagging where we want to assign category tag to the articles.

We propose a model on Urdu News text classification which achieve maximum accuracy that gives best result on our own created dataset by applying different machine learning algorithms and identify which algorithm is best to predict our dataset. Here we use eight pre-defined classes which are Sports, Business & Economics, and Science & Technology Entertainment news as an input into ten different machine learning technique. Our method contains five primary processes: stop words removal, stemming, feature vector, applying the machine learning algorithms and assign the class to the sentence.

1. **Methodology**

Our methodology contains a step-wise procedure; we started from the Urdu language corpus collection and then used some preprocessing techniques for features selection to apply actual

Classification algorithms. The flow chart in Fig-1 summarizes the process which we followed for our technique.

## **Flow Chart:**

Corpus collection

Input File

Stop word removal

Results

Classification

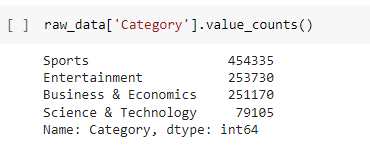
**Logistic Regression**

Feature Vector

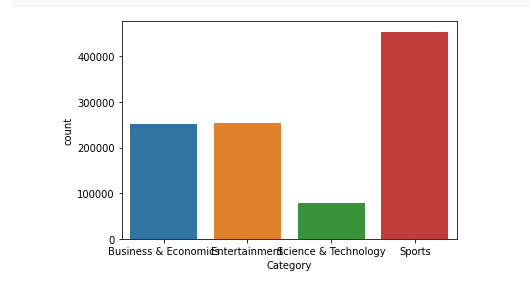
* 1. **Corpus Collection.**

Extensive training data plays a vital role in the development of a model that uses supervised learning algorithm. For this purpose, we write multiple crawlers to collect data from different news websites, e.g., bbcurdu.com, arynews and urdupoint.com. In total, we collected 141289 words. Data is collected category-wise in the text files, and categories are as follow: Business & Economics, Entertainment, sports, Science & Technology.

### Fig : 1.1



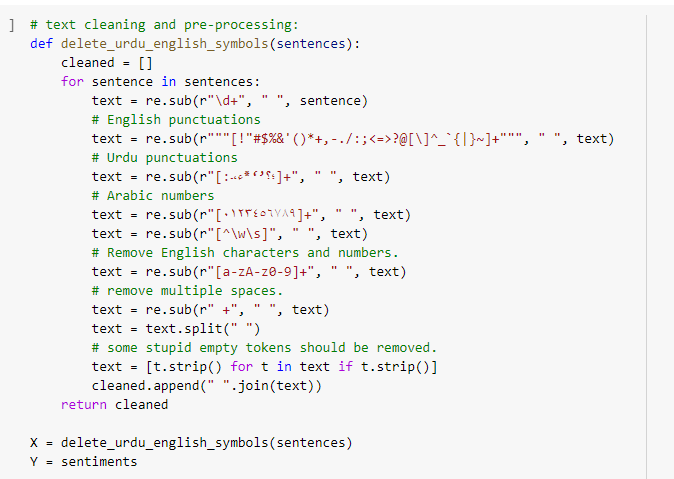
### Fig: 1.2



* 1. **Stop Words Removal.**

The words which are either not useful for the proposed classification models or used as prepositions are included in the stop words list. In our case, we maintained a list of stop words to omit from our text to extract meaningful data for the classifiers.

### Fig: 2.1



## **Feature Selection.**

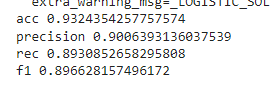
Feature selection is an important part of building machine learning models. We will be using the chi square test of independence to identify the important features.

1. **Results**

After all preprocessing techniques, we then implement our dataset on machine learning algorithm. We divide our dataset into training 80% and testing 20%.

Logistic regression was predicted as the best solution for this type of classification the output given by the model is stated below (Accuracy, precision, Recall, F1 score).

## Fig: 3.1



1. **CONCLUSION**

We downloaded our dataset from Mendeley website on Urdu news heading. After collecting our dataset, we pass it on preprocessing techniques to filter our data. Then finally we apply different machine learning algorithm to train our Urdu dataset. We found Logistic Regression is best algorithm for text classification that gives almost 93% accuracy to predict our class.

1. **References**
2. [**https://www.kaggle.com/adnanzaidi/ml-urdu-hindi-news-headline-classification**](https://www.kaggle.com/adnanzaidi/ml-urdu-hindi-news-headline-classification)
3. [**https://www.koreascience.or.kr/article/JAKO202121055546977.pdf**](https://www.koreascience.or.kr/article/JAKO202121055546977.pdf)
4. [**https://www.researchgate.net/publication/339029343\_UrduHindi\_News\_Headline\_Text\_Classification\_by\_Using\_Different\_Machine\_Learning\_Algorithms**](https://www.researchgate.net/publication/339029343_UrduHindi_News_Headline_Text_Classification_by_Using_Different_Machine_Learning_Algorithms)