# Urdu News Classification using Application of Machine Learning Algorithms on News Headline

**Muhammad Badruddin Khan**

Information Systems Department
College of Computer and Information Sciences
Imam Mohammad ibn Saud Islamic University (IMSIU),
Riyadh, KSA

**Abstract**
Our modern 'information-hungry' age demands delivery of information at unprecedented fast rates. Timely delivery of noteworthy information about recent events can help people from different segments of life in number of ways. As world has become global village, the flow of news in terms of volume and speed demands involvement of machines to help humans to handle the enormous data. News are presented to public in forms of video, audio, image and text. News text available on internet is a source of knowledge for billions of internet users. Urdu language is spoken and understood by millions of people from Indian subcontinent. Availability of online Urdu news enable this branch of humanity to improve their understandings of the world and make their decisions. This paper uses available online Urdu news data to train machines to automatically categorize provided news. Various machine learning algorithms were used on news headline for training purpose and the results demonstrate that Bernoulli Naïve Bayes (Bernoulli NB) and Multinomial Naïve Bayes (Multinomial NB) algorithm outperformed other algorithms in terms of all performance parameters. The maximum level of accuracy achieved for the dataset was 94.278% by multinomial NB classifier followed by Bernoulli NB classifier with accuracy of 94.274% when Urdu stop words were removed from dataset. The results suggest that short text of headlines of news can be used as an input for text categorization process.

***Key words:***
*Text categorization, Machine learning, Naïve Bayes, Support vector machine, Logistic regression, Word Cloud, Urdu language*

## 1. Introduction

Historically newspaper with their textual news assisted by some images were major source of information for general public. Local, national and International news have different impact on minds of people. Different types of news are preferred by different segments of society. Youths may be interested in entertainment news whereas an investor may be concerned about news related to economy. Many people plan based on weather news and predictions. Politicians build their narratives after their detailed analysis of available news.

The advent of digital age resulted in new ways of transmission of ideas and information to public. Before internet, radio and TV were the main electronic medium to disseminate news. Internet allowed news to stay on their webpages and hence remain available to public to be read at any time. Since the digital text is searchable, web crawlers of Internet search engines made it possible to index the web thus enabling a person to reach or discover news available at different websites using some keywords.

With the emergence of Unicode, online text from different languages began to appear on numerous websites. Urdu is the first language of nearly 70 million people and is the second language of more than 100 million people, predominantly in Indian subcontinent [1]. The language has its prominence as the leading language for national discourse in Pakistan. Due to its importance, number of public service broadcaster like British Broadcasting Corporation [1] (BBC), Voice of America [2], Deutsche Welle [3] (Germany), NHK[4] (Japan) offer their news in Urdu language at their respective websites. Most of the Urdu newspapers have enhanced their world-wide readership by putting news content on their website. Hence there exist huge amount of news data in Urdu that can be used for machine learning purposes.

Machines cannot learn directly from huge amount of textual news data that is available at various news sites in unstructured form. It requires collection of news data in proper format followed by application of specific preprocessing methods to produce suitable input for machine learning algorithms. The application of machine learning algorithms on transformed data enables automatic extraction of interesting information from news data. The process to derive interesting information from textual data is called text mining. According to [2], text mining involves *"the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources."*. Thus text mining directly or indirectly assists in data-driven decision making. Text mining solutions for different areas including but not limited to customer care service, contextual

[1] https://www.bbc.com/urdu
[2] https://www.urduvoa.com/
[3] https://www.dw.com/ur/
[4] https://www3.nhk.or.jp/nhkworld/ur/news/

Manuscript received February 5, 2021
Manuscript revised February 20, 2021
https://doi.org/10.22937/IJCSNS.2021.21.2.27

advertising, cybercrime prevention have successfully demonstrated that the machines after learning can help humans to save cost and time to perform a task successfully. Automatic text categorization or text classification answers the challenge of organizing unstructured data that is growing on internet at unprecedented pace. Urdu news classification has received comparatively less attention in the academic world. This paper reports and compares performance of different machine learning algorithms while classifying Urdu news into four classes or categories. Details of the performed work are discussed in subsequent sections.

In Section 2 of this paper the related works are described. The section is followed by the portrayal of the text corpora in section 3 which also contains non-technical description of few machine learning algorithms that were used in this work. Section 4 describes the experiments and their brief interpretations and is followed by section 5 that concludes the research paper.

## 2. Related work

Text categorization (TC) has received attention of researchers for their investigation on resourceful language like English for news classification purpose. However, very few works can be found in literature for automatic Urdu news categorization.

In one of the pioneer work in automatic English text categorization [3] used Reuters-21578 and OHSUMED corpus to categorize English text using support vector machine. [4] compared efficiency of five machine learning algorithms in domain of text categorization with respect to different parameters. The role of preprocessing in improving the performance of machine learning algorithm was discussed by [5]. [6] discussed why naïve Bayes algorithm did not perform well for text categorization and proposed two empirical heuristics that made naïve Bayes performance comparable to support vector machine. Different research has applied different machine learning algorithm under different parameter settings and has reported their results. For example [7] have reported good results with their robust algorithm that categorize text more effectively and efficiently irrespective of language and domain of the text.[8] grouped wide range of works done in field of text categorization into three basic fields namely conventional methods, fuzzy logic based methods and deep learning-based methods.

Urdu, Persian and Arabic belongs to family of languages which are written from right to left. The processing techniques that can be applied on one language can also be useful for other language too. [9] presented results of experiments in which different classification data mining algorithms (C4.5, PART, RIPPER, OneRule) were applied on Arabic dataset. Number of classes were 6 in the dataset and no algorithm was able to give accuracy of more than 70 percentages. [10] handled dataset of 1000 documents with 5 classes and reported accuracy of 98.20% by SVM classifier after light stemming of corpus.

As compared to English, not too many attempts were made in Urdu language to face the challenge of text categorization. [11] presented SVM-based framework for classification of Urdu news headline. After performing stemming operation on their dataset, the accuracy achieved was 86.66%. In [12], the researchers reported 91.4% accuracy after application of multi-layered perceptron (MLP) on their developed Urdu news dataset COUNT19. [13] scrapped Urdu news headline from different web resources and developed their dataset with 10 categories. The best result as reported by them was upto 87% accuracy achieved by application of Ridge classifier.

## 3. Text Corpus and its Processing

### 3.1 Text Corpus

The text corpus used for the purpose of research work was downloaded from [14]. The dataset comprises of Urdu news from four distinct categories or classes. The four categories are Business & Economics, Science & Technology, Entertainment, and Sports. Total number of Urdu news in the subset of downloaded corpus after some editing due to some technical reasons was 111,859. Bar chart in Figure 1 presents number of Urdu news belonging to each category.

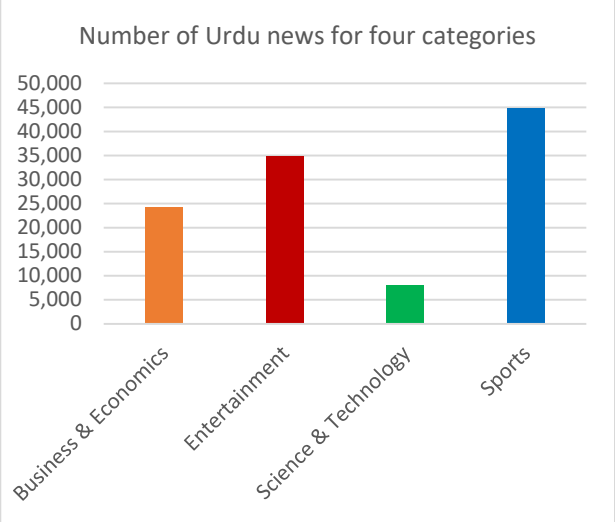


**Figure 1:** Number of Urdu news for four categories in the text corpus

Although the content of a news is the best candidate to be selected as attribute to categorize news however due to limited available computational capacity, this work used headline of news to develop classifiers. According to Wikipedia, purpose of a headline is to "quickly and briefly draw attention to the story". The headline describes "the nature of the article below it"[15]. It is naturally expected that the headlines should use keywords or tags that are suitable not only for the news but also for the categories of the news. Visualization of key tags for each category of news headings can help in better understanding of text corpus used in the research work.

Word cloud or tag cloud is a visualization to depict tags in given text [16]. Font size or color indicates importance of tag[17]. Figure 2, 3, 4 and 5 represents word clouds for the four categories. All

word clouds were developed in Python environment using Python wordcloud project[18]. Languages like Urdu and Arabic require their text to be transform in such a way that code from wordcloud project can use it as input. Both languages scripts have following two similar properties:
1) Scripts are written from right to left.
2) Based on the surrounding characters, the characters change their shape.

"Python Arabic Reshaper" [19] is a project to reconstruct Arabic sentences so that they can be used in applications that don't support Arabic script. The code from this project was used to transform Urdu news headline in order to construct word clouds in Urdu language.

**Figure 2**:Word cloud representing important tags from headings of Urdu news from Business & Economics category

**Figure 3**:Word cloud representing important tags from headings of Urdu news from Entertainment category

**Figure 4:** Word cloud representing important tags from headings of Urdu news from Science & Technology category

**Figure 5**:Word cloud representing important tags from headings of Urdu news from Sports category

In order to further understand the dataset, top 5 unigrams and bigrams were discovered using chi square. Table 1, 2, 3 and 4 represents the top 5 unigrams and bigrams for each of four categories.

| Category: Business & Economics | | | |
|---|---|---|---|
| Non-TF-IDF matrix with Token occurrence count | | TF-IDF matrix with Token importance | |
| Unigrams | Bigrams | Unigrams | Bigrams |
| مارکیٹ | قیمت میں | مارکیٹ | مارکیٹ میں |
| اضافہ | کی قیمتوں | روپے | کی قیمتوں |
| کمی | قیمتوں میں | کمی | قیمتوں میں |
| اسٹاک | کی قیمت | اضافہ | کی قیمت |
| روپے | پاکستان اسٹاک | اسٹاک | پاکستان اسٹاک |

**Table 1:** Top 5 unigrams and bigrams for Business & Economics category

It can be seen from table 1 that although top 5 bigrams are same for TF-IDF and non TF-IDF matrices, unigrams are different for two matrices. However, it is not necessary that bigrams importance of two matrices will always be same. In table 2, one can see the bigrams are also different in terms of importance. Since stop words were not removed from corpus, therefore some stop words can be seen in bigrams.

| Category: Entertainment | | | |
|---|---|---|---|
| Non-TF-IDF matrix with Token occurrence count | | TF-IDF matrix with Token importance | |
| Unigrams | Bigrams | Unigrams | Bigrams |
| ہالی | سلمان خان | ریلیز | کا ٹریلر |
| پاکستان | شاہ رخ | ہالی | شاہ رخ |
| وڈ | ہالی وڈ | وڈ | ہالی وڈ |
| ٹریلر | ہالی ووڈ | ٹریلر | ہالی ووڈ |
| فلم | ٹریلر جاری | فلم | ٹریلر جاری |

**Table 2:** Top 5 unigrams and bigrams for Entertainment category

| Category: Science & Technology | | | |
|---|---|---|---|
| Non-TF-IDF matrix with Token occurrence count | | TF-IDF matrix with Token importance | |
| Unigrams | Bigrams | Unigrams | Bigrams |
| ایپ | ئی فون | ایپ | ئی فون |
| متعارف | اسمارٹ فون | متعارف | سام سنگ |
| فیس | سام سنگ | فیس | اسمارٹ فون |
| بک | واٹس ایپ | بک | واٹس ایپ |
| فون | فیس بک | فون | فیس بک |

**Table 3:** Top 5 unigrams and bigrams for Science & Technology category

| Category: Sports | | | |
|---|---|---|---|
| Non-TF-IDF matrix with Token occurrence count | | TF-IDF matrix with Token importance | |
| Unigrams | Bigrams | Unigrams | Bigrams |
| شکست | پی سی | شکست | سری لنکا |
| کپ | ایس ایل | کپ | ایس ایل |
| کرکٹ | سی بی | کرکٹ | سی بی |
| ٹیسٹ | ون ڈے | ٹیسٹ | ون ڈے |
| ٹیم | ورلڈ کپ | ٹیم | ورلڈ کپ |

**Table 4:** Top 5 unigrams and bigrams for Sports category

Figures of word clouds and tables comprising chi-squared based important unigrams and bigrams help us in understanding the content of Urdu news headline. For example, important unigrams and bigrams of sports category tell about the sports that gets most attention of Urdu media i.e. Cricket.

In the next subsection, process flow for Urdu news classification will be described

## 3.2 Process flow for Urdu news classification

Urdu news headlines went through preprocessing stage in which corpus went through certain operations based on the requirement of the work. Since role of removal of stop words on classification performance was to be observed, therefore in one series of experiments, stop words were not removed in preprocessing and in other series, they were removed. Urdu stop word list was downloaded from [20]. Since a news headline is a short text therefore role of formulation of TF-IDF matrix was also planned to be observed.

Term frequency (TF) of a term *t* in document *d* is obtained by finding the frequency *t* in *d*. So,

$$\mathrm{TF}(t,d) = f(t,d) \qquad (1)$$

Inverse Document Frequency (IDF) is the ratio of total number of documents to the number of documents containing the term *t*. Mathematically,

$$idf(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|} \qquad (2)$$

where, $N$ is the total number of documents and $|\{d \in D: t \in d\}|$ is the number of documents where t appears. TF-IDF is then given by:

$$tf(t,d) \times idf(t,d) \qquad (3)$$

In one series of experiments, TF-IDF matrix was made and was used as input by different machine learning algorithms but in other series of experiments, simple matrix having occurrence count for every token was used. Both unigrams and bigrams tokens were extracted as features.

## 3.3 Machine learning algorithms

In this section, brief non-technical description of those machine learning algorithms in context of text categorization is provided that were used in this research work.

**Naïve Bayes algorithm:**

Naïve Bayes algorithm is a simple and efficient algorithm that has been reported to be widely used for text categorization task[21], [22]. It is reported to provide competitive classification performance when compared with other classifiers in domain of text categorization [23], [24]. Bernoulli naive Bayes (BNB) and multinomial naive Bayes (MNB) are two types of Naïve Bayes classifiers that were created when Bernoulli and Multinomial distribution models were incorporated into Bayesian framework. MNB usually outperformed then BNB in extensive experiments on real-life benchmarks with large vocabulary size as reported by [25].

**Support Vector Machine:**

[26] has discussed why SVM suits the task of text categorization in the light of properties of text. Text tends to have high dimensional input space and SVM classifiers have potential to handle such large feature space. Moreover, most text categorization problems are linear separable and SVMs also help in finding such linear separators. [26] also mentions that SVMs eliminate the requirement for feature selection because of their ability to generalize in high dimensional space. The robustness of the algorithm can be seen by promising results in all experiments.

**Logistic regression:**

Due to its close relationship with SVM as mentioned by [27], logistic regression, got attention of machine learning community. Logistic regression has been successfully used in text categorization problem.

## 4. Experiments and Results

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. The experiments were performed using 5-fold cross validation on training data that means training data was further divided into training and validation set. The experiments were completed using Python using implementations of different algorithms provided by Scikit-Learn library[28].

**Experiments without removal of Urdu stop words**:

In this category of experiments, extraction of features from the text of Urdu news headlines was performed *without* pre-processing step of removal of stop words thus transforming headlines into feature vectors. Different classifiers were constructed using 80% of training data using with setting of 5-fold cross-fold validation. The resultant classifier was tested on 20% testing data.

Both Bernoulli Naïve Bayes and multinomial Naïve Bayes algorithms are proven to be good in NLP-related tasks therefore both of them were used in Urdu news classification task. Sklearn's MultinomialNB and BernoulliNB are implementations of Naive Bayes that are built with NLP-related tasks in mind.


**Figure 6:** Confusion matrix for Bernoulli Naive Bayes Classifier

Different performance parameters' values for the above experiment is given in table 5.

| Category of news | Accuracy in percentage | Precision | Recall | F1-score |
|---|---|---|---|---|
| Business & Economics | 93.867% | 0.94 | 0.96 | 0.95 |
| Entertainment | | 0.93 | 0.93 | 0.93 |
| Science & Technology | | 0.8 | 0.87 | 0.83 |
| Sports | | 0.98 | 0.95 | 0.96 |

**Table 5**: Different performance parameters' values for Bernoulli Naïve Bayes classifier with non TF-IDF setting

Table 6 shows different performance parameter for Multinomial Naïve Bayes. Since news headlines are short unstructured text, Multinomial Naïve Bayes in this scenario was unable to outperform Bernoulli Naïve Bayes algorithm.

| Category of news | Accuracy in percentage | Precision | Recall | F1-score |
|---|---|---|---|---|
| Business & Economics | 93.786% | 0.94 | 0.96 | 0.95 |
| Entertainment | | 0.93 | 0.92 | 0.92 |
| Science & Technology | | 0.77 | 0.88 | 0.83 |
| Sports | | 0.98 | 0.95 | 0.96 |

**Table 6:** Different performance parameters' values for Multinomial Naïve Bayes classifier with non TF-IDF setting

With the same setting, when SVM and logistic regression algorithms were applied on the training dataset, performance in terms of accuracy were slightly low as can be seen in table 7 and 8.

| Category of news | Accuracy in percentage | Precision | Recall | F1-score |
|---|---|---|---|---|
| Business & Economics | 92.834% | 0.93 | 0.95 | 0.94 |
| Entertainment | | 0.89 | 0.94 | 0.91 |
| Science & Technology | | 0.87 | 0.72 | 0.79 |
| Sports | | 0.97 | 0.95 | 0.96 |

**Table 7:** Different performance parameters' values for SVM classifier with non TF-IDF setting

| Category of news | Accuracy in percentage | Precision | Recall | F1-score |
|---|---|---|---|---|
| Business & Economics | 91.963% | 0.93 | 0.93 | 0.93 |
| Entertainment | | 0.88 | 0.91 | 0.9 |
| Science & Technology | | 0.84 | 0.8 | 0.82 |
| Sports | | 0.96 | 0.94 | 0.95 |

**Table 8:** Different performance parameters' values for Logistic regression classifier with non TF-IDF setting

To avoid the issue of number of occurrence issue that can vary in small and big document, "Term Frequency times Inverse Document Frequency" is used. In the next experiment, the simple intuition was tested that TF-IDF should not play big role in improvement of performance of the classifier since the size of headline is same for small or big news. Results described in table 9 proves the intuition.

| Category of news | Accuracy in percentage | Precision | Recall | F1-score |
|---|---|---|---|---|
| Business & Economics | 93.661% | 0.93 | 0.96 | 0.94 |
| Entertainment | | 0.92 | 0.92 | 0.92 |
| Science & Technology | | 0.79 | 0.87 | 0.82 |
| Sports | | 0.97 | 0.95 | 0.96 |

**Table 9:** Different performance parameters' values for Bernoulli Naïve Bayes classifier with TF-IDF setting

For multinomial Naïve Bayes classifier, simple intuition is that the results should not be much different since the news headline length is very small. Table 10 testifies this intuition.

| Category of news | Accuracy in percentage | Precision | Recall | F1-score |
|---|---|---|---|---|
| Business & Economics | 93.679% | 0.94 | 0.96 | 0.95 |
| Entertainment | | 0.92 | 0.92 | 0.92 |
| Science & Technology | | 0.82 | 0.81 | 0.81 |
| Sports | | 0.97 | 0.96 | 0.96 |

**Table 10:** Different performance parameters' values for Multinomial Naïve Bayes classifier with TF-IDF setting

Table 11 and 12 represent the performance of support vector machine (SVM) and logistic regression classifiers.

| | Accuracy in percentage | Precision | Recall | F1-score |
|---|---|---|---|---|
| Business & Economics | 89.723% | 0.92 | 0.89 | 0.9 |
| Entertainment | | 0.87 | 0.9 | 0.89 |
| Science & Technology | | 0.85 | 0.58 | 0.69 |
| Sports | | 0.91 | 0.95 | 0.93 |

**Table 7:** Different performance parameters' values for Support vector machine classifier with TF-IDF setting

It can be seen that SVM classifier performance slightly deteriorated in TF-IDF setting as compared to table 7 where SVM reached almost 93% accuracy.

| Category of news | Accuracy in percentage | Precision | Recall | F1-score |
|---|---|---|---|---|
| Business & Economics | 91.475% | 0.93 | 0.92 | 0.93 |
| Entertainment | | 0.87 | 0.92 | 0.89 |
| Science & Technology | | 0.86 | 0.75 | 0.8 |
| Sports | | 0.95 | 0.93 | 0.94 |

**Table 12:** Different performance parameters' values for Logistic regression classifier with TF-IDF setting

The summarized picture of the series of experiments without removal of Urdu stop words is given in table 13.

| | Non TF-IDF setting | | | TF-IDF setting | | |
|---|---|---|---|---|---|---|
| Classifiers | Accuracy | Precision | Recall | Accuracy | Precision | Recall |
| Bernoulli NB | 93.87% | 0.91 | 0.93 | 93.66% | 0.9 | 0.92 |
| Multinomial NB | 93.79% | 0.91 | 0.93 | 93.68% | 0.91 | 0.91 |
| SVM | 92.83% | 0.92 | 0.89 | 89.72% | 0.89 | 0.83 |
| Logistic regression | 91.96% | 0.9 | 0.9 | 91.48% | 0.9 | 0.88 |

**Table 13:** Comparison of four machine learning algorithm performance with stop words present as tokens

**Discussion:** Even though stop words were not removed, the accuracy of Bernoulli NB and Multinomial NB classifiers reached more than 93% in Non TF-IDF setting. It should be noted that in TF-IDF setting, performance of Bernoulli NB, Multinomial NB and logistic regression classifiers decreased slightly however SVM classifier performance saw a notable decline. In terms of average precision and recall, SVM performed worst in TF-IDF setting with precision and recall declining to 0.89 and 0.83

**Experiments after removal of Urdu stop words**:

Table 14,15,16 and 17 present the results of experiment for the four classifiers with Urdu stop words **removed** in pre-processing stage and the matrix of tokens has number of occurrence instead of TF-IDF as value.

| Category of news | Accuracy in percentage | Precision | Recall | F1-score |
|---|---|---|---|---|
| Business & Economics | 94.274% | 0.95 | 0.96 | 0.95 |
| Entertainment | | 0.93 | 0.93 | 0.93 |
| Science & Technology | | 0.81 | 0.88 | 0.84 |
| Sports | | 0.98 | 0.95 | 0.97 |

**Table 14:** Different performance parameters' values for Bernoulli Naïve Bayes classifier with non TF-IDF setting

| Category of news | Accuracy in percentage | Precision | Recall | F1-score |
|---|---|---|---|---|
| Business & Economics | 94.278% | 0.95 | 0.96 | 0.95 |
| Entertainment | | 0.93 | 0.93 | 0.93 |
| Science & Technology | | 0.8 | 0.89 | 0.84 |
| Sports | | 0.98 | 0.96 | 0.97 |

**Table 15:** Different performance parameters' values for Multinomial Naïve Bayes classifier with non TF-IDF setting

| Category of news | Accuracy in percentage | Precision | Recall | F1-score |
|---|---|---|---|---|
| Business & Economics | 93.380% | 0.94 | 0.95 | 0.95 |
| Entertainment | | 0.89 | 0.94 | 0.92 |
| Science & Technology | | 0.88 | 0.76 | 0.82 |
| Sports | | 0.97 | 0.95 | 0.96 |

**Table 16:** Different performance parameters' values for Support vector machine classifier with non TF-IDF setting

| Category of news | Accuracy in percentage | Precision | Recall | F1-score |
|---|---|---|---|---|
| Business & Economics | 90.425% | 0.93 | 0.91 | 0.92 |
| Entertainment | | 0.84 | 0.92 | 0.88 |
| Science & Technology | | 0.86 | 0.76 | 0.81 |
| Sports | | 0.96 | 0.91 | 0.93 |

**Table 17:** Different performance parameters' values for Logistic regression classifier with non TF-IDF setting

Table 18, 19, 20 and 21 present the results of experiment for the four classifiers with Urdu stop words **removed** in pre-processing stage and the matrix of tokens has value of TF-IDF for every token. Table 22 compares the performance of four machine learning algorithms in terms of accuracy, precision and recall with stop words removed from dataset in pre-processing stage.

| Category of news | Accuracy in percentage | Precision | Recall | F1-score |
|---|---|---|---|---|
| Business & Economics | 93.661% | 0.93 | 0.96 | 0.94 |
| Entertainment | | 0.92 | 0.92 | 0.92 |
| Science & Technology | | 0.79 | 0.87 | 0.82 |
| Sports | | 0.97 | 0.95 | 0.96 |

**Table 18:** Different performance parameters' values for Bernoulli Naïve Bayes classifier with TF-IDF setting

| Category of news | Accuracy in percentage | Precision | Recall | F1-score |
|---|---|---|---|---|
| Business & Economics | 93.679% | 0.94 | 0.96 | 0.95 |
| Entertainment | | 0.92 | 0.92 | 0.92 |
| Science & Technology | | 0.82 | 0.81 | 0.81 |
| Sports | | 0.97 | 0.96 | 0.96 |

**Table 19:** Different performance parameters' values for Multinomial Naïve Bayes classifier with TF-IDF setting

| Category of news | Accuracy in percentage | Precision | Recall | F1-score |
|---|---|---|---|---|
| Business & Economics | 89.723% | 0.92 | 0.89 | 0.95 |
| Entertainment | | 0.87 | 0.9 | 0.92 |
| Science & Technology | | 0.85 | 0.58 | 0.82 |
| Sports | | 0.91 | 0.95 | 0.96 |

**Table 20:** Different performance parameters' values for Support vector machine classifier with TF-IDF setting

| Category of news | Accuracy in percentage | Precision | Recall | F1-score |
|---|---|---|---|---|
| Business & Economics | 91.475% | 0.93 | 0.92 | 0.93 |
| Entertainment | | 0.87 | 0.92 | 0.89 |
| Science & Technology | | 0.86 | 0.75 | 0.8 |
| Sports | | 0.95 | 0.93 | 0.94 |

**Table 21:** Different performance parameters' values for Logistic regression classifier with TF-IDF setting

| | Non TF-IDF setting | | | TF-IDF setting | | |
|---|---|---|---|---|---|---|
| Classifiers | Accuracy | Precision | Recall | Accuracy | Precision | Recall |
| Bernoulli NB | 94.27% | 0.92 | 0.93 | 93.66% | 0.9 | 0.92 |
| Multinomial NB | 94.28% | 0.91 | 0.93 | 93.68% | 0.91 | 0.91 |
| SVM | 93.38% | 0.92 | 0.9 | 89.72% | 0.89 | 0.83 |
| Logistic regression | 90.43% | 0.9 | 0.88 | 91.48% | 0.9 | 0.88 |

**Table 22:** Comparison of four machine learning algorithm performance with stop words present as tokens

**Discussion:** After removal of stop words, the accuracy of Bernoulli NB and Multinomial NB classifiers reached more than 94% in Non TF-IDF setting. It should be noted that in TF-IDF setting, performance of Bernoulli NB, Multinomial NB decreased slightly however SVM classifier performance saw a big dip whereas logistic regression classifier increased its accuracy slightly in TF-IDF setting. In terms of average precision and recall, SVM classifier performed worst in TF-IDF setting with precision and recall declining to 0.89 and 0.83 similar to result of experiment that was done without removal of stop words.

## 5. Conclusion

The huge unstructured data in the digital world, is one of the venue where the application of various machine learning techniques is possible. In this work, text categorization was performed on Urdu news headline using four machine learning algorithms namely Bernoulli NB, Multinomial NB, SVM and Logistic regression. Bernoulli NB and Multinomial NB provided robust performance with the accuracy value not fluctuating too much. Good results from the experiments show that news headlines can be trusted as reliable indicator to predict the category of news. This insight is important as headlines are small sentences and to work on them requires little computational resources.

The work can be extended further by using different available lexical databases like WordNet. By the combination of lexicons and machine learning techniques, categories can be discovered based on "meaning" of the news.

## References


[1] "Urdu language | History, Script, & Words," *Encyclopedia Britannica*. https://www.britannica.com/topic/Urdu-language (accessed Feb. 25, 2021).

[2] "Marti Hearst: What Is Text Mining?" https://people.ischool.berkeley.edu/~hearst/text-mining.html (accessed Feb. 25, 2021).

[3] T. Joachims, "Text Categorization with Support Vector Machines," *Proc Eur. Conf Mach. Learn. ECML98*, Jan. 1998, doi: 10.17877/DE290R-5097.

[4] S. Dumais, J. Platt, D. Heckerman, and M. Sahami, "Inductive learning algorithms and representations for text categorization," in *Proceedings of the seventh international conference on Information and knowledge management*, New York, NY, USA, Nov. 1998, pp. 148–155, doi: 10.1145/288627.288651.

[5] A. Basu, C. Watters, and M. Author, "Support Vector Machines for Text Categorization.," Jan. 2003, p. 103, doi: 10.1109/HICSS.2003.1174243.

[6] Sang-Bum Kim, Kyoung-Soo Han, Hae-Chang Rim, and Sung Hyon Myaeng, "Some Effective Techniques for Naive Bayes Text Classification," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 11, pp. 1457–1466, Nov. 2006, doi: 10.1109/TKDE.2006.180.

[7] E. S. Tellez, D. Moctezuma, S. Miranda-Jiménez, and M. Graff, "An automated text categorization framework based on hyperparameter optimization," *Knowl.-Based Syst.*, vol. 149, pp. 110–123, Jun. 2018, doi: 10.1016/j.knosys.2018.03.003.

[8] A. Dhar, H. Mukherjee, N. S. Dash, and K. Roy, "Text categorization: past and present," *Artif. Intell. Rev.*, Sep. 2020, doi: 10.1007/s10462-020-09919-1.

[9] M. Al-diabat, "Arabic text categorization using classification rule mining," *Appl. Math. Sci.*, vol. 6, no. 81, pp. 4033–4046, 2012.

[10] I. Hmeidi, M. Al-Ayyoub, N. Abdulla, A. Almodawar, R. Abooraig, and N. A. Ahmed, "Automatic Arabic text categorization: A comprehensive comparative study," *J. Inf. Sci.*, vol. 41, pp. 114–124, Jan. 2014, doi: 10.1177/0165551514558172.

[11] K. Ahmed, M. Ali, S. Khalid, and M. Kamran, "Framework for Urdu News Headlines Classification.," *J. Appl. Comput. Sci. Math.*, no. 21, 2016.

[12] S. A. Hamza, B. Tahir, and M. A. Mehmood, "Domain Identification of Urdu News Text," in *2019 22nd International Multitopic Conference (INMIC)*, Nov. 2019, pp. 1–7, doi: 10.1109/INMIC48123.2019.9022736.

[13] S. Hassan and A. Zaidi, "Urdu News Headline, Text Classification by Using Different Machine Learning Algorithms," Feb. 2019.

[14] K. Hussain, N. Mughal, I. Ali, S. Hassan, and S. M. Daudpota, "Urdu News Dataset 1M," vol. 3, Jan. 2021, doi: 10.17632/834vsxnb99.3.

[15] "Headline," *Wikipedia*. Feb. 11, 2021, Accessed: Feb. 25, 2021. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Headline&oldid=1006104286.

[16] "Tag cloud," *Wikipedia*. Jan. 23, 2021, Accessed: Feb. 25, 2021. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Tag_cloud&oldid=1002276039.

[17] "An Assessment of Tag Presentation Techniques." http://www2007.org/htmlposters/poster988/ (accessed Feb. 25, 2021).

[18] "WordCloud for Python documentation — wordcloud 1.8.1 documentation." https://amueller.github.io/word_cloud/ (accessed Feb. 26, 2021).

[19] "arabic_reshaper - crates.io: Rust Package Registry." https://crates.io/crates/arabic_reshaper/0.1.4 (accessed Feb. 26, 2021).

[20] "Urdu Stopwords List." https://kaggle.com/rtatman/urdu-stopwords-list (accessed Mar. 01, 2021).

[21] D. D. Lewis, "Naive (Bayes) at forty: The independence assumption in information retrieval," in *Machine Learning: ECML-98*, Berlin, Heidelberg, 1998, pp. 4–15, doi: 10.1007/BFb0026666.

[22] Y. H. Li and A. K. Jain, "Classification of Text Documents," *Comput. J.*, vol. 41, no. 8, pp. 537–546, Jan. 1998, doi: 10.1093/comjnl/41.8.537.

[23] "An extensive empirical study of feature selection metrics for text classification | The Journal of Machine Learning Research." https://dl.acm.org/doi/10.5555/944919.944974 (accessed Feb. 28, 2021).

[24] "A Comparative Study on Feature Selection in Text Categorization | Proceedings of the Fourteenth International Conference on Machine Learning." https://dl.acm.org/doi/10.5555/645526.657137 (accessed Feb. 28, 2021).

[25] A. Mccallum and K. Nigam, "A Comparison of Event Models for Naive Bayes Text Classification," *Work Learn Text Categ*, vol. 752, May 2001.

[26] "Text categorization with Support Vector Machines: Learning with many relevant features | SpringerLink." https://link.springer.com/chapter/10.1007/BFb0026683 (accessed Feb. 28, 2021).

[27] V. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed. New York: Springer-Verlag, 2000.

[28] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. 85, pp. 2825–2830, 2011.

**Dr. Muhammad Badruddin Khan** obtained his doctorate in 2011 from Tokyo Institute of Technology, Japan. He is a full-time assistant professor in department of Information Systems of Al-Imam Muhammad Ibn Saud Islamic University since 2012. The research interests of Dr. Khan lie mainly in the field of data and text mining. He is currently involved in number of research projects related to machine learning and Arabic language including pandemics prediction, Arabic sentiment analysis, improvement of Arabic semantic resources, Stylometry, Arabic Chatbots, trend analysis using Arabic Wikipedia, Arabic proverbs classification, cyberbullying and fake content detection, and violent/non-violent video categorization using Youtube video content and Arabic comments, and has published number of research papers in various conferences and journals. He is also co-author of a book on machine learning.