Journal of Xi'an Shiyou University, Natural Science Edition ISSN: 1673-064X

Sentence-Level Classification of Web-Extracted Data in Urdu Language Text (ULT)

Somia Ali [1,*], Uzma Jamil [1], Mamoona Jabbar [1], Muhammad Assad Jabbar [1]

[1] Department of Computer Science, Government College University, Faisalabad, Pakistan

*

Abstract: The most prominent and dominant way of communication in the current digital era is text instead of using sound, emotions, pictures, and animation. Millions of users are using the internet because of its real-time availability. Social media is one of the most promising sources of information. On social media, the usage of local language is increasing day by day. People share their points of view on different topics of interest on social media. Natural Language Processing (NLP) is an emerging domain for the processing of different languages for different purposes. People from different cultures, interests, and knowledge areas share their ideas, opinions, and occasions like food festivals, sports, death and murder, politics, law and order, terrorist attacks, and others in the local language on social media. In this research, a sentence-level classification is performed for extracting the different occasions from social media. Those extracted occasions are then classified into different classes. For occasion classification, Machine learning (ML) classifiers are used. For the evolution of the proposed work, performance measuring parameters (precision, recall, F1-score, and accuracy) are used. In our experiment, linear SVC and Ridge classifier shows the best accuracy of 83%. In the future, Deep learning classifiers can be used to enhance the accuracy of text classification.

Keywords: Machine Learning; Classification; Urdu Language Text; Sentence-level classification-; Natural Language Processing

---

1. Introduction

Social media is dominating and a major source of communication in the current era. People share their thoughts, opinions, feelings, occasions, and marketing advertisements through social media. These factors generate heterogeneous data that causes challenges for the worthy extraction of sentiments [1], law and order prediction, risk factor analysis [2], construction of timeline, opinion mining, decision-making system, social media monitoring, spam detection, retrieval of information, classification of e-mail [3], classification of sentences [4], modeling topics [5], content labeling and trend findings. Natural Language Processing comes into the ground as a serious and effective developed area with the aim of different grammar rules, semantic rules for major languages and their structure, and syntactical and lexical constructs of different languages (multilingual). We have an enormous amount of data to process in the present years.

Urdu is also spoken in India, Iran, and Afghanistan. About 340 million people use Urdu on social media for various purposes. Text mining and text analysis are essentially two tasks of extracting information and text statistics. Machine learning has also recently been used to acquire and process large amounts of data. Thus, natural language processing (NLP) together with text analytics, machine learning, and data mining provide services for the creation and evaluation of large data sources. Text mining and text analysis are used as similar terms in the literature. Statistical, linguistic, and machine learning ideas used to build various models are trained on a dataset, and new results are computed from this training information.

On the other hand, text analysis uses the information obtained from the text mining process to create graphs and visualize graph data. The most visited website in Pakistan has its content in the Urdu language. Urdu presents many problems in natural language processing because Urdu uses formal and informal verb forms and masculine and feminine nouns. In Urdu, the text is arranged from right to left with different grammatical structures.

a) Subject-object-verb [6]
b) No letter capitalization
c) Diacritics

d) Free word order

The Urdu language consists of 38 basic characters that can be joined with different characters. The joined nature of Urdu language alphabets is called ligature in Urdu [7]. These characters are given in Figure 1.

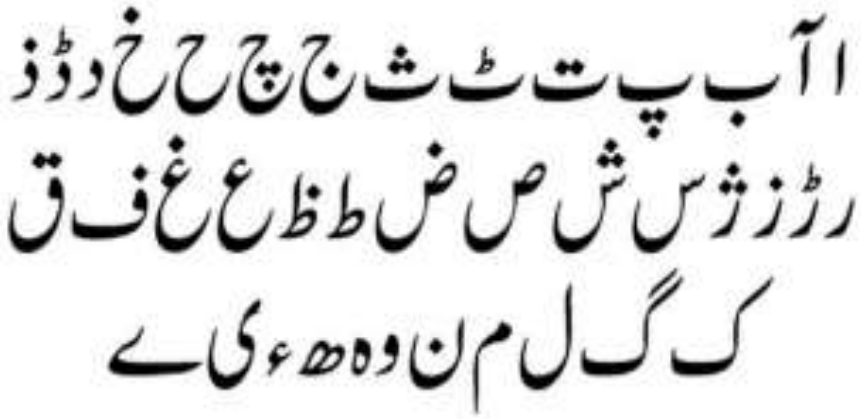

Figure 1. Basic Urdu characters [7]

These joining natures enrich Urdu with almost 24,000 ligatures. This alphabet set is being used as a superset for all Urdu script-based language alphabets like Arabic and Persian having 28 and 32 alphabets respectively [6]. An example of these ligatures is given in Table 1.

Table 1. Urdu words with ligatures and characters [7]

Urdu Words	Urdu Ligatures	Urdu Characters
فیصل آباد	فیصل ، آباد	ف، ی، ص، ل، آ، ب، ا، د
نظارے	نظا،ر،ے	ن ، ظ ، ا ، ر ، ے
سیاست	سیا ، ست	س ، ی ، ا ، س ، ت

There are still gaps in processing resources such as part-of-speech (PoS) markers, titles, entity identifiers, and annotation tools for sentence detection and classification in Urdu. Many people are not familiar with the meaning and usage of some Urdu words. This creates linguistically ambiguous content, making sentence classification a difficult and counterintuitive task. The lack of sufficient resources/datasets is another significant challenge for data-driven and knowledge-based sentence extraction and classification methods.

The distributed representation of words on vector spaces is called word embedding. In this method, language words are represented on multiple vector spaces [8]. An example is given in Figure 2. A sentence can be classified as a specific situation, action, or happening that occurs in a certain period. The information that is extracted can be of different types like literacy ratio, top trends, famous foods, and some religious sentences, etc.

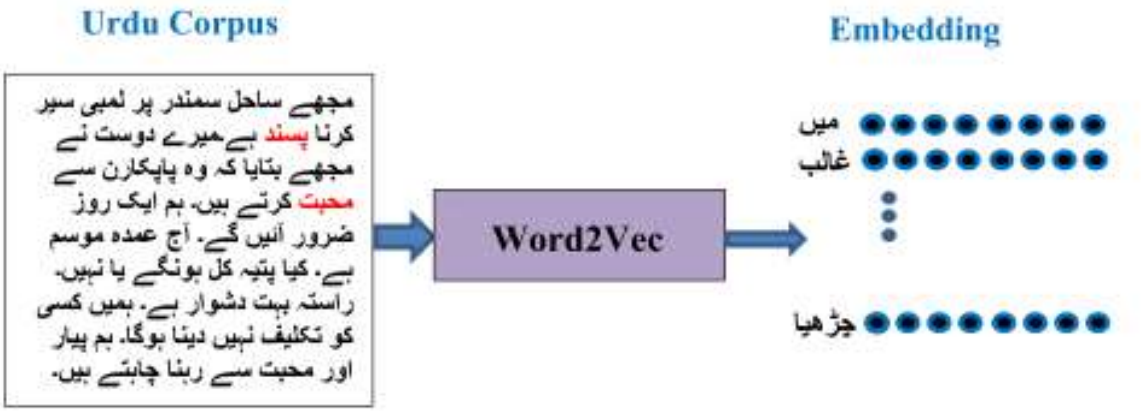

Figure 2. Vector representation of Urdu text [8]

To address this knowledge area, a suitable classification technique will be performed for sentence classification in this research. Furthermore, unstructured data will also be transformed into different sentence classes. There is a big gap in sentence classification in the Urdu language due to the limited set of data and poor training in deep learning and machine learning models. As the dataset is limited so the training process becomes ambiguous and gives less accuracy in the classification of sentences. To address this problem, a deep learning approach will be used for Urdu Language Text (ULT) on a web-extracted dataset for sentence-level classification.

In the past, the Urdu language had limited processing resources as the dataset is limited. Annotators, part-of-speech (PoS) taggers, and translators were also limited for this language. So less work is done in the past by researchers. Feature-based classification for Urdu text documents has been using machine learning models for the last few years. To the best of my knowledge, many classification models are not evaluated yet. Different machine learning and deep learning models were used for Urdu text classification with some limitations.

Research Question 1: How significantly the deep learning classifier will categorize the occasions into their appropriate classes?

Objective 1: Selection and implementation of suitable deep learning classifier to categorize the occasions into their appropriate classes.

The aim is to classify sentences with a deep learning model using the larger dataset of Urdu Language Text (ULT). As the dataset is in larger amounts, the training process will be effective and testing gives better results for Urdu Language Text (ULT) extracted from the web. Due to the limited resources in Urdu text, there is a huge gap to cover in different domains like classification with unstructured datasets, a multiclass classification that helps to easily classify different kinds of occasions at different levels like document level, sentence level, phrase level, etc. There is a dire need to overcome these gaps to help the predictors easily analyze Urdu text.

- Desired to provide services in this problem domain
- An interesting and challenging task in classification is to classify sentences of Urdu language text.
- Many deep learning approaches give limited accuracy for Urdu Language Text (a resource-poor language) in sentence classification.
- Many occasions were not classified in the past due to a limited amount of data.

2. Related Work

The Urdu language has a complex and rich morphological script but it is a resource-poor language [9]. This fact makes it a challenging task to process text automatically. Also, the long text document dataset is unavailable and become a major obstacle for the Urdu language TDC. The researchers design a dataset for this purpose to make it easy. Then a Single-layer Multi-size Filters Convolutional Neural Network (SMFCNN) for the classification of the dataset. The performance of this network is analyzed n different data split ratios. One novel model for the classification of short text based on event-level information on the dataset extracted from text is explained in [10]. This information is used for deep neural networks as

supplementary knowledge. The importance of this supplementary knowledge is measured with an attention mechanism.

A char-based model is selected for Chinese word segmentation. Then classification is performed with is information. The result shows that the given method outperforms the state-of-art methods. The work presented herein [11] is inspired by work done in other languages for fake news detection. Ensemble learning models are explored here for the improvement of fake news detection in the Urdu Language. An annotated corpus is created manually here from Urdu news articles. Three famous machine learning classifiers as Naïve Bayes (NB), SVM, and Decision Tree (DT) are used for the classification of fake news. On the other hand, five ensemble models are used including voting, grading, stacking, ensemble selection, and cascade generalization. Text categorization is performed here on Urdu language news headlines by using four classifiers of machine learning [12]. Those machine learning algorithms are Multinomial NB, Logistic regression, Bernoulli NB, and SVM. The results show that Bernoulli NB and Multinomial NB give a robust performance with minimum accuracy fluctuation. Those good results show that the headlines of news are a great source of information and can be used for the prediction of news categories. For the distribution of Urdu words, skip-grams are used in the word2vec model [13]. Word embedding is used for the numerical representation of words. For unlabeled datasets, word embedding is used for the representation of semantic and syntactic meanings. For NLP and computers linguistic word embedding generation is considered the most powerful tool for processing. This word embedding tool is used for sentiment analysis, machine translation, semantic analysis, information retrieval, transliterations information retrieval, etc.

Work done for hate speech detection is from the Twitter dataset [14]. A large corpus is prepared for hate speech detection with sentiments and aspect-level annotations for the Urdu language with the help of experts in the team. Data evaluation and analysis are performed. After that machine learning algorithms are used to train a classifier. The major focus is to analyze the most occurring problems as high dimensional feature vectors, highly skewed classes, and highly sparse data representation. For the huge amount of data for text analysis topic modeling is considered as a backbone [15]. For the retrieval of semantic information, topic modeling is used. Different parsing techniques are used here for the extraction of datasets from different sites as HTML and XML datasets [15]. Preprocessing is applied for data cleaning. Then for the classification purpose, some machine learning classifiers are used as Naïve Bayes and Logistic Regression. Three different models are used for topic modeling Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), and Probabilistic Latent Semantic Analysis (PLSA). Urdu is a resource-poor language. It has inadequate resources and a complicated and complex morphological script that is a hurdle in text processing [16]. There is no large and free dataset available for text processing in Urdu. This is considered a big issue nowadays. To address this issue, a large dataset is developed with Urdu news documents. For the first time, hierarchical text classification is performed for the Urdu language by using a deep neural network. HMLSTM is proposed and used here for this purpose. This model can predict each level category and classify the categories hierarchically.

A large number of tweets are collected for the creation of the dataset due to the unavailability of the Urdu dataset for many experiments. Then many variants are generated based on the common word co-occurrence from the dataset. Then many n-grams are extracted from the dataset by developing a topic modeling approach based on multiple feature representations. Hashtags are analyzed for the comprehension of topics.

Table 2 shows the summary of literature from different aspects of ML and DL models.

Table 2. Summary of existing literature

Sr. No #	Year of Publication	Result	References
1	2020	In this research, the authors present SMFCNN and compare the performance with different machine learning classifiers.	[17]
2	2020	Event detection is performed on short-text Chinese datasets with the help of conceptual information extracted from the dataset.	[10]
3	2020	Word embedding is a powerful tool for the processing of natural language. It provides the semantic and syntactic processing of language.	[13]
4	2021	This article shows that news headlines are a great source of information that can be used for the prediction of fake news or categorization of news.	[12]
5	2021	This search is performed for the improvement of Urdu language fake news detection by using three machine learning (ML) classifiers and five ensemble learning models.	[9]
6	2021	A new huge dataset is created for the detection of hate speech from Twitter. The three major problems are resolved here for the improvement of the detection process.	[14]
7	2021	The work is done on the classification of text and topic modeling by extracting the dataset from different sites. Two machine learning classifiers ate used for classification purposes. For topic modeling, three LSA, PLSA, and LDA are used.	[15]

8	2021	For text classification, a hierarchical approach is used by creating a large dataset of Urdu news and then applying the deep neural network for classification purposes. A model is developed here named Hierarchical Multi-layer LSTMs (HMLSTM)	[16]
9	2021	A dataset is produced from Urdu tweets due to the unavailability of a large dataset. The LDA and NMF algorithms are analyzed for topic modeling by automatic labeling.	[19]
10	2021	Four different resource-rich datasets are used for text processing. These datasets are translated into low-resource language by using translators.	[18]
11	2021	Contextualized text representation is used for fake news detection with deep neural classification. Three different classifiers (MLP, SLP, and CNN) are used for implementation. Gaussian Noise layer is added to a contextualized text representation.	[20]
12	2021	Fake news regarding COVID-19 is detected. Preprocessing techniques are used.TF-IDF is used for data representation. Eight different machine learning algorithms and four different deep learning models are used.	[21]
13	2021	Six different emotions are classified using machine learning classifiers. A dataset is generated by extracting different emotions from social media platforms like Facebook, Instagram, YouTube, and Twitter.	[22]
14	2021	Urdu Text Sentiment Analysis (UTSA) is performed here for the analysis of sentiments. Sentiments are classified into positive and negative classes. Different deep-learning methods are used for classification.	[23]

15	2021	The dataset is gathered from six different universities in the province of Sindh. The sample of 273 students is generated from the dataset. The students belong to the age group of 18-25 years.	[24]

3. Dataset

A large dataset is created for occasion classification. The dataset contains 1624 labeled sentences. The dataset is collected from different social media sites and newsgroups for different occasions. The web crawler is used for collecting the dataset which is a PHP-based web scraper. The diversity of the sentences gives us multi-classes. This dataset contains 14 types of occasions. The subset of a dataset can be useful for other researchers.

4. Pre-processing

Some preprocessing techniques were used for preparing the dataset for ML classifiers. The techniques i.e. noise removal technique and sentence labeling were used here for preparing the dataset. All other language words were removed from the dataset. Hyperlinks, URLs, and all special symbols were also removed.

4.1. Annotation

For annotation of a dataset, one M.Phil. (Urdu) level language expert was engaged. He analyzes the sentences and assigns the labels for respective classes.

a) Assign a class label to each sentence

b) Removal of ambiguous sentences

c) For similar sentences, one class is assigned

Assign one of 14 occasions i.e. terrorist attack, national, sports, entertainment, politics, safety, earthquakes, farad and corruption, sexual assault, weather, accidents, forces, inflation, murder, and death to each sentence. Occasions and their labels are shown in Table 4.

4.2. Stop Words

In the next step, stop words were removed by filtering the sentences and tokenization of words. All unnecessary words that do not contribute to classification were removed from the dataset. Some tokenized words are shown in Table 3. An example of text before and after stop words is shown in Figure 3.

	Sentences	Labels	text_cleaned
0	یہ بہت خوشی کی بات ہے کہ وہ کراچی آئے ہیں ک...	1	[یہ, بہت, خوشی, کی, بات, ہے, کہ, وہ, کراچی, آئ...]
1	یہ محض مودی کو خوش کرنے کے لیے کیا گیا	1	[یہ, محض, مودی, کو, خوش, کرنے, کے, لیے, کیا, گیا]
2	... ان کا کہنا تھا کہ وہ نریندر مودی کے اس فیصلے	1	[... ,ان, کا, کہنا, تھا, کہ, وہ, نریندر, مودی, کے]
3	کیا وہ بھی اس خوشی میں شامل ہوں گے	1	[کیا, وہ, بھی, اس, خوشی, میں, شامل, ہوں, گے]
4	...تاہم خوش قسمتی سے اب تک یہ احتجاج یا انقلاب پ	1	[...,تاہم, خوش, قسمتی, سے, اب, تک, یہ, احتجاج, یا]

Figure 3. Stop words removal

4.3. Boundary

In this step, we decide the length of the sentence to a specific number. In the dataset, sentences vary in length from 5 to 250 words. We limit the sentence to 150 words. After 150 words, a new sentence was created from the given sentence. Some simple sentences example is shown in Figure 4. Figure 5 shows an example of sentence boundary setting.

Table 3. Sentence tokenization

Urdu Sentence	Tokens
ٹرمپ کے بیان کے بعد ریکارڈ کی تصیح کیلئے	ٹ ر م پ ک ے ب ی ا ن ک ے ب ع د ر ی ک ا ر ڈ ک ی ت ص ی ح ک ی ل ئ
ایشین ہاکی کپ کے فائنل میچ میں بھارت کے ساتھ	ای ش ی ن ہ اک ی ک پ ک ے ف ائ ن ل م ی چ م ی ں ب ھ ارت ک ے س ات ھ

Table 4. Labels of occasions

Occasion	Label
Terrorist attack	0
National,	1
Sports,	2
Entertainment,	3
Politics,	4
Safety,	5
Earthquakes,	6
Farad and corruption,	7
Sexual assault,	8
Weather,	9
Accidents,	10
Forces	11
Inflation	12
Murder and death	13

Figure 6 shows the occasion that is classified in this experiment. Each occasion is assigned a label starting from 0 and ending at 13. There are a total of 14 occasions that are classified in this research. In Figure 7, each occasion is shown with the number of sentences in that category.

	class	details
0	دہشت گرد حملہ	کراچی : ایس ایچ او گڈاپ شاہ رخ پاور فل آفیسر
1	قومی	...پاکستان کا یمن کے معاملے پر موثر کردار ادا کر
2	قومی	...سابق وزیراعظم یوسف رضاگیلانی کے بھانجے کی گاڑ
3	قومی	...علیم خان کی جائیدادیں ملک سے باہر ، تفصیلات آ
4	قومی	...سرکاری افسر اکرام نوید سے50کروڑ چرائے اور شہب

Figure 4. Some sentences examples

	Sentences	Labels	text_cleaned	Lemmatized
0	...یہ بہت خوشی کی بات ہے کہ وہ کراچی آئے ہیں ک	1	...یہ, بہت, خوشی, کی, بات, ہے, کہ, وہ, کراچی, آئ]	...میں, بہت, خوشی, کہ, بات, ہونا, کہنا, میں, کرا]
1	یہ محض مودی کو خوش کرنے کے لیے کیا گیا	1	[یہ, محض, مودی, کو, خوش, کرنے, کے, لیے, کیا, گیا]	...میں, محض, مودی, کو, خوش, کرنا, کہ, لینا, کیا]
2	... ان کا کہنا تھا کہ وہ نریندر مودی کے اس فیصلے	1	... ,ان, کا, کہنا, تھا, کہ, وہ, نریندر, مودی, کے]	... ,میں, کا, کہنا, تھا, کہنا, میں, نریندر, مودی]
3	کیا وہ بھی اس خوشی میں شامل ہوں گے	1	[کیا, وہ, بھی, اس, خوشی, میں, شامل, ہوں, گے]	[کیا, میں, بھی, میں, خوشی, میں, شامل, ہونا, گے]
4	...ذہم خوش قسمتی سے اب تک یہ احتجاج یا انقلاب پ	1	...,ذہم, خوش, قسمتی, سے, اب, تک, یہ, احتجاج, یا]	... ,ذہم, خوش, قسمتی, سے, اب, تکنا, میں, احتجاج]

Figure 5. Setting the boundary

5. Experiment

Various traditional machine learning classifiers were performed in this experiment. The basic aim for performing many classifiers is to find the most accurate and efficient result/accuracy for the classification of multi-classes for imbalance datasets for Urdu language text.

5.1. Correlation

The chi-square technique was used for finding the correlation between words and labels. Examples are shown in table 5. Unigrams and bigrams tokens were used for creating feature space from those features. 1559 features were used here in this experiment. On the other hand, one-hot encoding and word embedding were also used in our feature space to enhance the feature space.

5.2. Feature Vector Generating Techniques

In this vector generation, the text is represented in numerical form. This is the type of input that a machine-learning model used for classification. There are many feature vector-generating techniques used for processing text. The following techniques are used in this research.

5.2.1. Word Embedding Models

This represents each word in the numerical text that is considered a feature vector for ML classifiers. This creates real values of dens vector which captures the semantic, contextual, and syntactical meaning of the word. It also assigns related weighted values to similar words. TF-IDF and One-hot encoding is used here [26].

5.2.2. One Hot Encoding

ML classifiers cannot deal with textual data directly. It should be converted into real values. In Tables 6 and 7, an example is shown for a sentence with real values.

	class	category_id
0	دہشت گرد حملہ	0
1	قومی	1
517	کھیل	2
729	تفریح	3
908	سیاست	4
928	امن و امان	5
952	زلزلہ	6
967	فراڈ اور کرپشن	7
983	جنسی حملہ	8
1005	موسم	9
1305	حادثات	10
1455	افواج	11
1605	مہنگائی	12
1617	قتل اور موت	13

Figure 6. Classes with categories

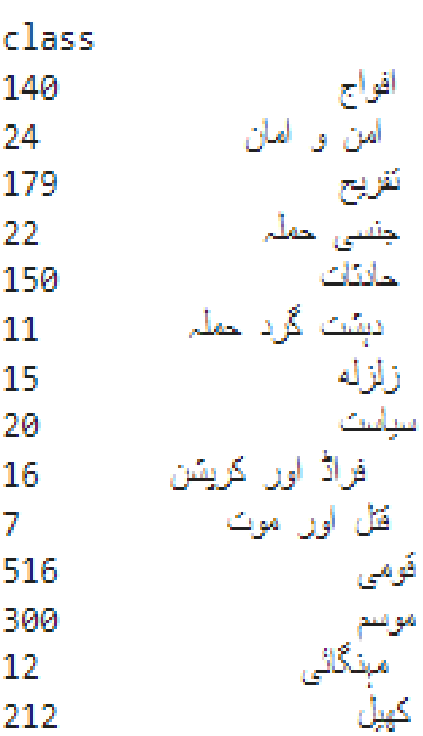

Figure 7. Number of sentences for each class

5.2.3. TF-IDF

This is a feature engineering technique that is used to transform textual data into real values. This is one of the famous methods used for generating feature vectors for textual datasets.

Table 5. Unigram and bigram with chi-square

Tag	Most Correlated unigram	Most correlated bigram
افواج	فوج آر جی	پاک فوج ایس پی پی آر
امن و امان	مصر جنگ ہلاک	سری لنکن موبائل فون کام شروع
تفریح	دیپکا اداکارہ شادی	سری لنکن موبائل فون کام شروع
جنسی حملہ	انڈو امریکا فلسطینی	ڈونلڈ ٹرمپ سعودی عرب امریکی ریاست
حادثات	روڈ ٹریفک حادثہ	مسافر بس الم ناک ٹریفک حادثہ
دہشت گرد حملہ	دورہ پاور سینٹر	پاک بحریہ دہشت گردی پاک چین
زلزلہ	جہاز تباہ انڈونیشیا	سری لنکن امریکی ریاست مسافر طیارہ
سیاست	ماں اسرائیل اسرائیلی	سعودی عرب اسرائیلی وزیراعظم افراد ہلاک

سمندر ڈوب گر تباہ قسط نمبر	نیشنل کھل گاہ	فراڈ اور کرپشن
پیغام جاری ویڈیو پیغام مطالبہ کردیا	صارفین کشمیری تشدد	قتل اور موت
سپریم کورٹ وزیراعظم عمران عمران خان	حکومت عمران وزیراعظم	قومی
علاقوں بارش دھار بارش موسلا دھار	موسم کراچی بارش	موسم
ڈالر 4 روپے روپے اوپن اوپن مارکیٹ	اوپن انٹربینک مارکیٹ	مہنگائی
شعیب ملک ایس ایل نیوزی لینڈ	میچ نیوزی ٹیم	کھیل

Table 6. Occasion sentences

Urdu Sentence
کاشف ہاکی کھیلتا ہے
چھ ستمبر کو یوم دفاع ہے

Table 7. One-hot encoding

Sentence	کھیلتا	ہاکی	کاشف	دفاع	یوم	ستمبر	چھ
1	1	1	1	0	0	0	0
2	0	0	0	1	1	1	1

The flow of the research work is given in Figure 8. Textual data in the Urdu language is used for input. Data preprocessing is performed for the removal of noisy data. For feature selection, feature engineering techniques are used. After feature engineering, ML classifiers are implemented on the dataset with training and testing splits of 70%:30% respectively.

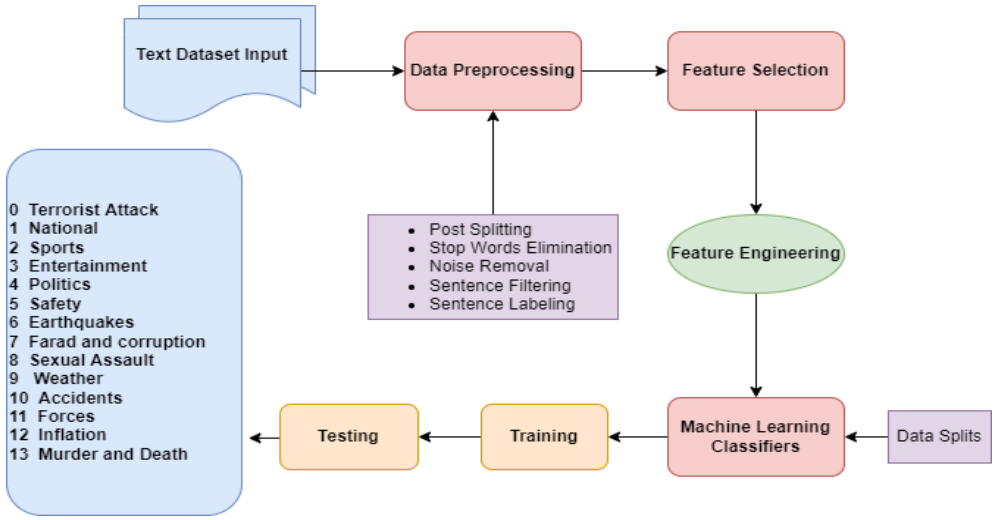

Figure 8. The flow of research work

6. Performance Measuring Parameters

In this research, performance measuring parameters i.e. precision, recall, F1-score, and accuracy are used to evaluate the results. This is decided due to the multi-class classification and also due to the unevenness of the dataset.

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$F1 = \frac{2 * (Precision * Recall)}{(Precision + Recall)} \quad (3)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

Where,

TP = True Positive

TN = True Negative

FP = False Positive

FN = False Negative

Where precision is the state or degree of being precise. In other words, we can say that precision is the number of significant digits to which a number is calculated on can be measured reliably. While the recall is the ratio of true positive instances and the sum of true positive and false negative instances in experimental work. F1-score shows the harmonic mean of precision and recall. It is used to rate the performance of classifiers.

7. Results and Discussions

7.1. Machine Learning Classifiers

Eight different ML classifiers are used for the experiment.

7.1.1 Random-Forest (RF)

Many decision trees are comprised in this classifier. It shows very less accuracy on our dataset. Its accuracy is 81% with a depth of 3 and random states of 1 value. The summary is shown in Table 8.

Table 8. Results for Random-Forest (RF) classifier

Labels	Precision	Recall	F1-Score	Support
0	0.09	0.08	0.08	2
1	0.84	0.92	0.88	163
2	0.78	0.78	0.78	80
3	0.88	0.73	0.80	59
4	0.04	0.11	0.21	6
5	0.05	0.11	0.22	7
6	0.12	0.23	0.25	7
7	0.90	0.92	0.91	5
8	0.25	0.17	0.20	6
9	0.93	0.98	0.95	107

10	0.60	0.84	0.70	44
11	0.80	0.76	0.78	42
12	1.00	0.40	0.57	5
13	0.78	0.78	0.78	3
Accuracy			0.81	536
Macro-Avg	0.43	0.40	0.40	536
Weighted-Avg	0.78	0.81	0.79	536

7.1.2. K-Neighbors

It measures the similarity distance between the nearest neighbors. In our experiment, the value of neighbors=10 with leaf size=30. This model shows an accuracy of 80%. The summary is shown in Table 9.

Table 9. Results for K-Neighbor (KN) classifier

Labels	Precision	Recall	F1-Score	Support
0	0.33	0.50	0.40	2
1	0.81	0.87	0.84	163
2	0.72	0.86	0.78	80
3	0.79	0.63	0.70	59
4	0.11	0.12	0.13	6
5	0.32	0.41	0.39	7
6	0.22	0.23	0.23	7
7	0.33	0.32	0.38	5
8	0.22	0.22	0.22	6
9	0.91	0.97	0.94	107
10	0.69	0.93	0.80	44
11	0.89	0.76	0.82	42
12	1.00	0.40	0.57	5
13	0.76	0.78	0.77	3
Accuracy			0.80	536
Macro-Avg	0.44	0.42	0.42	536
Weighted-Avg	0.76	0.80	0.77	536

7.1.3. Nearest Centroid

It is also called the nearest prototype. It assigns labels to the dataset. In our experiment, we use the Euclidean metric with no threshold. It shows an accuracy of 76%. The summary is shown in Table 10.

Table 10. Results for Nearest Centroid (NC) classifier

Labels	Precision	Recall	F1-Score	Support
0	0.33	0.50	0.40	2
1	0.67	0.91	0.77	163
2	0.93	0.70	0.80	80
3	0.85	0.68	0.75	59
4	0.14	0.17	0.15	6
5	0.20	0.14	0.17	7
6	0.25	0.14	0.18	7
7	0.33	0.36	0.35	5
8	0.26	0.28	0.27	6
9	0.88	0.93	0.90	107

10	1.00	0.70	0.83	44
11	0.85	0.67	0.75	42
12	1.00	0.40	0.57	5
13	0.75	0.73	0.74	3
Accuracy			0.76	536
Macro-Avg	0.51	0.42	0.45	536
Weighted-Avg	0.78	0.76	0.76	536

7.1.4. Logistic Regression

It is also a supervised learning classifier that is used to predict the probability of the target variable. In our experiment, we use zero random states. This model shows an accuracy of 78%. The summary is shown in Table 11.

Table 11. Results for Logistic Regression (LR) classifier

Labels	Precision	Recall	F1-Score	Support
0	0.23	0.22	0.23	2
1	0.60	0.96	0.74	163
2	0.95	0.61	0.74	80
3	0.95	0.61	0.74	59
4	0.42	0.39	0.41	6
5	0.23	0.23	0.23	7
6	0.12	0.16	0.14	7
7	0.32	0.34	0.33	5
8	0.56	0.56	0.56	6
9	0.95	0.99	0.97	107
10	0.97	0.82	0.89	44
11	0.85	0.69	0.76	42
12	0.76	0.76	0.76	5
13	0.78	0.79	0.78	3
Accuracy			0.78	536
Macro-Avg	0.38	0.34	0.35	536
Weighted-Avg	0.77	0.78	0.75	536

7.1.5. Multinomial Naïve Bayes (MNB)

This model normally requires integer feature counts. But in our experiment, we use fractional feature count such as TF-IDF. This model shows an accuracy of 77%. The summary is shown in Table 12.

Table 12. Results for Multinomial Naive Bayes (MNB) classifier

Labels	Precision	Recall	F1-Score	Support
0	0.45	0.46	0.45	2
1	0.62	0.96	0.75	163
2	0.95	0.66	0.78	80
3	0.87	0.58	0.69	59
4	0.22	0.22	0.22	6
5	0.34	0.43	0.42	7
6	0.68	0.67	0.68	7
7	0.12	0.17	0.15	5

8	0.22	0.22	0.22	6
9	0.91	0.99	0.95	107
10	0.92	0.80	0.85	44
11	0.84	0.64	0.73	42
12	0.78	0.76	0.77	5
13	0.67	0.69	0.68	3
Accuracy			0.77	536
Macro-Avg	0.36	0.33	0.34	536
Weighted-Avg	0.75	0.77	0.74	536

7.1.6. Perceptron

It is a linear classifier that performs certain computations for detecting features. In our experiment with alpha value = 0.0001 and max_iter= 50 and with validation_fraction= 0.1. These models show an accuracy of 80%. The summary is shown in Table 13.

Table 13. Results for Perceptron classifier

Labels	Precision	Recall	F1-Score	Support
0	0.20	0.50	0.29	2
1	0.84	0.88	0.86	163
2	0.87	0.75	0.81	80
3	0.74	0.81	0.77	59
4	0.21	0.17	0.18	6
5	0.20	0.14	0.17	7
6	0.22	0.18	0.20	7
7	0.73	0.79	0.78	5
8	0.09	0.17	0.12	6
9	0.93	0.98	0.95	107
10	0.95	0.82	0.88	44
11	0.88	0.71	0.79	42
12	0.60	0.60	0.60	5
13	0.78	0.76	0.77	3
Accuracy			0.80	536
Macro-Avg	0.45	0.45	0.44	536
Weighted-Avg	0.81	0.80	0.80	536

7.1.7. Linear SVC

This model provides the best fitting hyperplane depending upon the data we provide which categorizes the dataset. In this experiment, the value of C=1.0 with max_iter=1000. This model shows an accuracy of 83%. The summary is shown in Table 14.

Table 14. Results for Linear SVC Classifier

Labels	Precision	Recall	F1-Score	Support
0	0.25	0.50	0.33	2
1	0.79	0.94	0.86	163
2	0.88	0.80	0.84	80
3	0.83	0.75	0.79	59
4	0.24	0.51	0.35	6
5	0.49	0.21	0.32	7

6	0.50	0.14	0.22	7
7	0.66	0.58	0.63	5
8	0.22	0.17	0.19	6
9	0.94	0.99	0.96	107
10	0.93	0.91	0.92	44
11	0.87	0.79	0.82	42
12	0.67	0.40	0.50	5
13	0.22	0.21	0.22	3
Accuracy			0.83	536
Macro-Avg	0.48	0.44	0.45	536
Weighted-Avg	0.81	0.83	0.81	536

7.1.8. Ridge Classifier

This classifier is based on the regression method which converts the labeled data into digits and solves a problem with the regression method. In this experiment, alpha= 1.0 with solver='sag'. This model shows an accuracy of 83%. The summary is shown in Table 15

Table 15. Results for Ridge classifier

Labels	Precision	Recall	F1-Score	Support
0	0.43	0.38	0.41	2
1	0.76	0.94	0.84	163
2	0.89	0.79	0.83	80
3	0.83	0.75	0.79	59
4	0.45	0.27	0.38	6
5	0.49	0.18	0.25	7
6	0.50	0.14	0.22	7
7	0.43	0.27	0.35	5
8	0.56	0.29	0.39	6
9	0.95	0.99	0.97	107
10	0.93	0.91	0.92	44
11	0.85	0.81	0.83	42
12	0.67	0.40	0.50	5
13	0.72	0.67	0.71	3
Accuracy			0.83	536
Macro-Avg	0.46	0.41	0.42	536
Weighted-Avg	0.80	0.83	0.81	536

Each model result is captured in a box plot. Figure 9 shows the results of every model.

8. Comparison of proposed work with previous research

The work done here in this research is a classification of 14 occasions based on their sentences. Our SVC model gives 83% accuracy. The previous work [27] classifies 12 events with different machine-learning classifiers. Random forest and Linear Regression give 80% accuracy. Our proposed model is performing best and giving the outclass results as compared to other reported results as shown in figure 10 below.

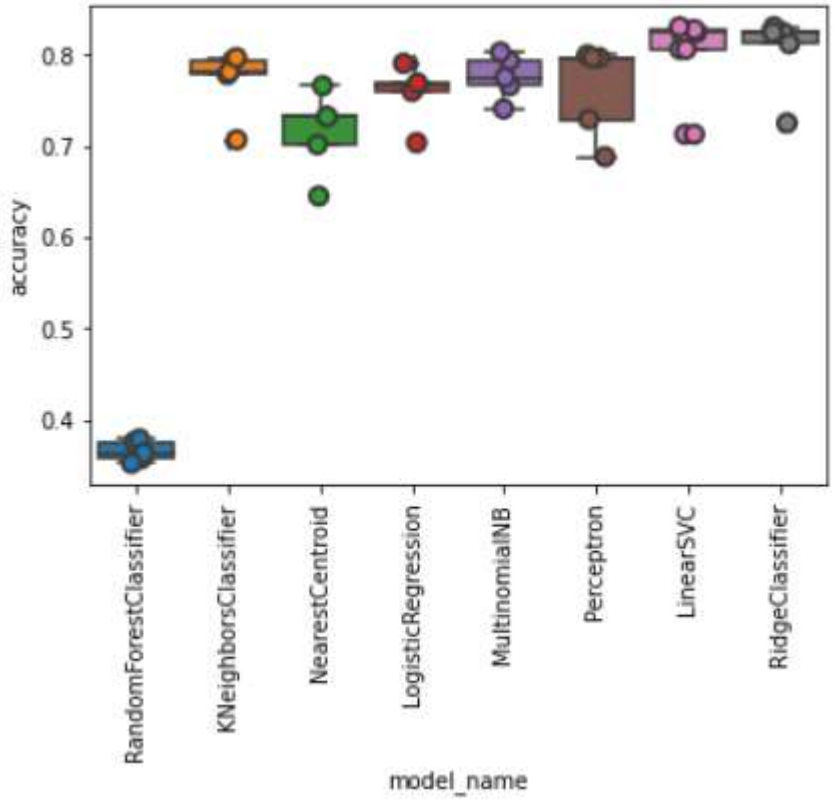

Figure 9: Results of models

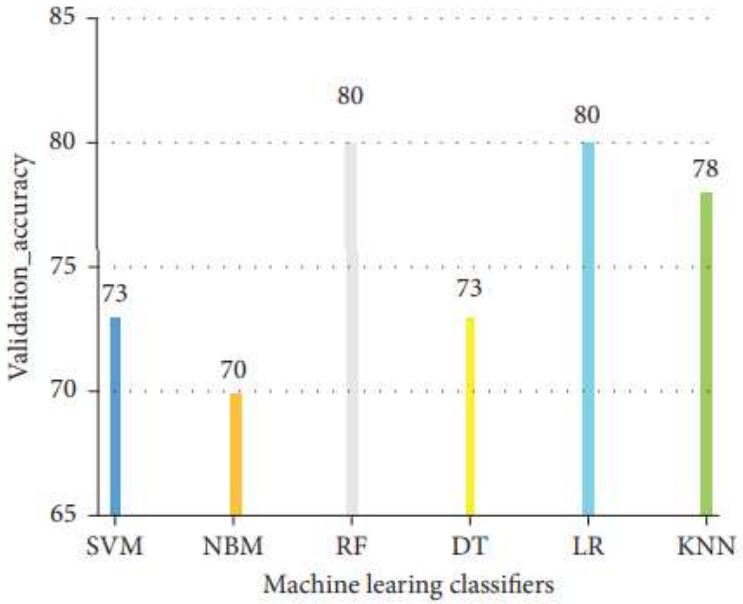

Figure 10: Machine learning algorithms' accuracy [27]

9. Conclusion

The absence of a sufficient amount of resources is a significant obstacle to Urdu language text research. In this research sentence classification is performed. For feature vector generation, one-hot encoding and TF-IDF models are used here because these models give better results for textual data. There is a need to enhance the word-embedding models that can be used for Urdu text on a huge corpus. Nonetheless, among all the referenced feature-vector strategies, TF-IDF beat them all. It showed the most noteworthy exactness. At this level, the classification of multi-class occasions is performed on uneven datasets. Due to the unevenness of the dataset, model performance is affected. In this research, TF-IDF and one-hot-encoding are better feature-generation methods when compared with previous Urdu language text word-embedding

models. An uneven number of occurrences in the dataset affected the general accuracy. In our experiment, linear SVC and Ridge classifier shows better accuracy (83%) when comparing them with other models' accuracies. In the future, we can work on a dataset by balancing it. By balancing the dataset, the performance of models can be improved. We can also use deep learning models in the future.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sailunaz, K,; Alhajj, R.; Emotion and sentiment analysis from Twitter text. Journal of Computational Science. 2019 Sep 1;36:101003.
2. Capet, P.; Delavallade, T.; Nakamura, T.; Sandor, A.; Tarsitano, C.; Voyatzi, S.; A risk assessment system with automatic extraction of event types. InInternational Conference on Intelligent Information Processing 2008 Oct 19 (pp. 220-229). Springer, Boston, MA.
3. Deerwester, S.; Dumais, S. T.; Furnas, G. W.; Landauer, T. K.; Harshman, R.; Indexing by latent semantic analysis. Journal of the American society for information science. 1990 Sep;41(6):391-407.
4. Mikolov, T.; Le, Q. V.; Sutskever, I.; Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168. 2013 Sep 17.
5. Alghamdi, R.; Alfalqi, K.; A survey of topic modeling in text mining. Int. J. Adv. Comput. Sci. Appl.(IJACSA). 2015 Jan;6(1).
6. Daud, A.; Khan, W.; Che, D.; Urdu language processing: a survey. Artificial Intelligence Review. 2017 Mar;47(3):279-311.
7. Pal, U.; Sarkar, A.; Recognition of printed Urdu script. InSeventh International Conference on Document Analysis and Recognition, 2003. Proceedings. 2003 Aug 1 (Vol. 3, pp. 1183-1183). IEEE Computer Society.
8. Kumhar, S. H.; Kirmani, M. M.; Sheetlani, J.; Hassan, M.; Word embedding generation for Urdu language using Word2vec model. Materials Today: Proceedings. 2021 Jan 25.
9. Akhter, M. P.; Jiangbin, Z.; Naqvi, I. R.; Abdelmajeed, M.; Mehmood, A.; Sadiq, M. T.; Document-level text classification using single-layer multisize filters convolutional neural network. IEEE Access. 2020 Feb 27;8:42689-707.
10. Yin, W.; Shen, L.; A short text classification approach with event detection and conceptual information. InProceedings of the 2020 5th International Conference on Machine Learning Technologies 2020 Jun 19 (pp. 129-135).
11. Akhter, M. P.; Zheng, J.; Afzal, F.; Lin, H.; Riaz, S.; Mehmood, A.; Supervised ensemble learning methods towards automatically filtering Urdu fake news within social media. PeerJ Computer Science. 2021 Mar 9;7:e425.
12. Khan, M. B.; Urdu News Classification using Application of Machine Learning Algorithms on News Headline. International Journal of Computer Science & Network Security. 2021;21(2):229-37.
13. Kumhar, S. H.; Kirmani, M. M.; Sheetlani, J.; Hassan, M.; Word embedding generation for Urdu language using Word2vec model. Materials Today: Proceedings. 2021 Jan 25.
14. Ali, M. Z.; Rauf, S.; Javed, K.; Hussain, S.; Improving hate speech detection of Urdu tweets using sentiment analysis. IEEE Access. 2021 Jun 9;9:84296-305.
15. Kumar, N.; Suman, R. R.; Kumar, S.; Text Classification and Topic Modelling of Web Extracted Data. In2021 2nd Global Conference for Advancement in Technology (GCAT) 2021 Oct 1 (pp. 1-8). IEEE.
16. Javed, T. A.; Shahzad, W.;, Arshad, U.; Hierarchical Text Classification of Urdu News using Deep Neural Network. arXiv preprint arXiv:2107.03141. 2021 Jul 7.
17. Akhter, M. P.; Jiangbin, Z.; Naqvi, I. R.; Abdelmajeed, M.; Mehmood, A.; Sadiq, M. T.; Document-level text classification using single-layer multisize filters convolutional neural network. IEEE Access. 2020 Feb 27;8:42689-707.
18. Ghafoor, A.; Imran, A. S.; Daudpota, S. M.; Kastrati, Z.; Batra, R.; Wani, M. A.; The Impact of Translating Resource-Rich Datasets to Low-Resource Languages Through Multi-Lingual Text Processing. IEEE Access. 2021 Sep 3;9:124478-90.
19. Latif, S.; Shafait, F.; Latif, R.; Analyzing LDA and NMF Topic Models for Urdu Tweets via Automatic Labeling. IEEE Access. 2021 Sep 14;9:127531-47.
20. Samadi, M.; Mousavian, M.; Momtazi, S.; Deep contextualized text representation and learning for fake news detection. Information Processing & Management. 2021 Nov 1;58(6):102723.
21. Bangyal, W. H.; Qasim, R.; Ahmad, Z.; Dar, H.; Rukhsar, L.; Aman, Z.; Ahmad, J.; Detection of fake news text classification on COVID-19 using deep learning approaches. Computational and mathematical methods in medicine. 2021 Nov 15;2021.
22. Nabeel, Z.; Mehmood, M.; Baqir, A.; Amjad, A.; Classifying emotions in roman urdu posts using machine learning. In2021

Mohammad Ali Jinnah University International Conference on Computing (MAJICC) 2021 Jul 15 (pp. 1-7). IEEE.

23. Naqvi, U.; Majid, A.; Abbas, S. A.; UTSA: Urdu text sentiment analysis using deep learning methods. IEEE Access. 2021 Aug 12;9:114085-94.

24. Saleem, S.; Khan, N. F.; Zafar, S.; Prevalence of cyberbullying victimization among Pakistani Youth. Technology in Society. 2021 May 1;65:101577.

25. Ahmed, K.; Ali, M.; Khalid, S.; Kamran, M.; Framework for Urdu News Headlines Classification. Journal of Applied Computer Science & Mathematics. 2016 Jan 1(21).

26. Akhter, M. P.; Jiangbin, Z.; Naqvi, I. R.; Abdelmajeed, M.; Fayyaz, M.; Exploring deep learning approaches for Urdu text classification in product manufacturing. Enterprise Information Systems. 2022 Feb 1;16(2):223-48.

27. Ali, D; Missen, M.M.; Husnain, M.; Multiclass event classification from text. Scientific Programming. 2021 Jan 13;2021.