Jana Nexus: Journal of Computer Science

ISSN(o): Applied
ISSN(p): Applied
DOI:
Journal Homepage: jananexuscs.com

| RESEARCH ARTICLE

Paper Title:

QUANTITATIVE SENTIMENT MINING FOR ROMAN URDU LANGUAGE

Ahmed Raza[1], Mubashir Ubaid Ullah[2], Kainat Saleem[2], Zaman Aslam[3] and Irfan Khalid[3]

1.School of Systems and Technology, University of Management and Technology, Lahore 54770, Pakistan.
2. Department of Computer Science, COMSATS University Islamabad, Lahore Campus, Pakistan.
3. Faculty the Information Technology, The University of Lahore, Lahore Campus, Pakistan.

Corresponding Author: Author's Name, Ahmed Raza

| ABSTRACT

Sentiment mining is the natural language processing task that helps to classify a large amount of web-based data for people's opinion-making. Most of the research work has been carried out for resource-rich languages but less resource language like Roman Urdu needs a considerable effort. In this paper, we classify Roman Urdu data according to the sentiment (Positive/Nega tive) using numerous feature selection techniques.

| KEYWORDS

Sentiment Mining, Natural Language Processing, Machine Learning

| ARTICLE INFORMATION

Received: 6 September 2025 **ACCEPTED: 10 October 2025** **PUBLISHED: November 2025**

ABSTRACT:

Sentiment mining is the natural language processing task that helps to classify a large amount of web-based data for people's opinion-making. Most of the research work has been carried out for resource-rich languages but less resource language like Roman Urdu needs a considerable effort. In this paper, we classify Roman Urdu data according to the sentiment (Positive/Negative) using numerous feature selection techniques. Feature selection techniques include Chi-square, Mutual Information, select form model that are being evaluated on different n-gram variations on a dataset of 11k Roman Urdu reviews. The dataset contains sentences from different domains like (Drama, Sports, Politics, Food etc.) are classified according to its nature. The experiments were performed by applying renowned machine learning and neural network classifiers. Machine learning classifiers incorporate logistic regression, support vector machine, decision tree, random forest, multinomial naïve Bayes and multilayer perceptron whereas convolutional neural network, long-short term memory and bidirectional long-short term memory belong to the neural network. The results are evaluated for both character-level and word-level variations of n-gram using evaluation measure accuracy and f1-score. Due to the diverse range of spellings being used in Roman Urdu we have measured results for different variations. We have achieved the highest accuracy 91.8% & 91.7% f1-score for bi-lstm meanwhile for the character-level we have 83.91% accuracy and 90.51% f1-score on 4-gram variation. For word-level analysis, 83.73% accuracy and 90.42% f1-score have been achieved on 1–4-gram variation. These results outperformed the baseline results for the Roman Urdu classification which shows the impact of feature selection techniques for sentiment mining classification.

Introduction:-

Social sites (Facebook, Twitter, discussion forums etc.) are the broad platform where people share their reviews, thoughts, feelings and opinions. Sentiment Mining (SM) is the way of inquiring the polarity (Positive/Negative) of these reviews, emotions etc. The data on social sites is in huge number that it is impossible to analyze and classify the data. Extracting important information from the data and classify it accordingly is the sentiment mining process [1]. SM can be performed at three various levels: Document-level, Sentence-level and a particular aspect-level. Major research focuses on resource-rich languages like (Chinese, English). But a huge gap for resource-poor languages like (Urdu, Roman Urdu) is a motivation for our work. We have performed sentiment mining for Roman Urdu due to two primary reasons. The initial one is Urdu/Hindi, which has more than 500 million speakers and among the top three widely spoken languages. Secondly, people feel comfortable communicating in Indo-Pak using Latin script (means using Roman Urdu writing 26-English alphabet) rather than using their native language style of communication [2]. The core purpose of sentiment mining is to classify a text input/review in the following steps: feature Extraction and/or feature selection and then the classification of sentiment. Various feature selection techniques have been used for developing a suitable feature set for classification using machine learning, neural network techniques. A process of extracting textual words depicting some sentiment is a feature extraction technique [3]. A Well-known feature extraction technique, Term Frequency-Inverse Document Frequency (TF-IDF) has been used for extracting features that not only calculate the occurrence of features but also measure the important information provided by features. The process of selecting the most relevant and efficient features for accurate prediction is feature selection. Various techniques have been used includes chi-square, mutual information etc. for feature selection [4].

The classification of sentiments has been performed by using various machine learning and neural network classifiers. Machine learning techniques include Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), Multinomial Naïve Bayes (MNB) and Multi-Layer Perceptron (MLP). The neural network techniques include Convolutional Neural Network (CNN), Long-Short Term Memory (LSTM) and Bidirectional LSTM (Bi-LSTM). Feature selection techniques play a vital role in achieving the best results in our work evaluating different evaluation measures.

The major contributions of this work comprised as follow:

1. Exploring different feature selection techniques for Roman Urdu sentiment mining.
2. Comparing various word-level and character-level n-gram variations for classifying sentiment by applying feature selection techniques.
3. Dig out the most suitable classifier by applying a combination of machine learning classifiers and neural network techniques.

The remaining paper is structured as follows: Section 2 is about Literature Review. Section 3 includes architecture details and structural methodology for feature selection and extraction techniques. The experimental setup, results and discussion are included in Section 4. Lastly, we conclude our work and future work in Section 5.

Literature Review:-

Sentiment mining for a resource-poor language like Roman Urdu is a challenging task because it doesn't have any rules for classification. Sentiment mining and natural language processing explain the issues, techniques, and methods that are being used for this process. Different types of techniques like machine learning, lexicon-based, corpus-based and complex network-based have been used for sentiment mining [6]. Sentiment mining techniques have been performed on three levels: Document-level, sentence-level and aspect-level. At the document-level, a document as a whole is held into consideration for classification of opinion. A sentence is either subjective type or objective type in a document but in the classification of document objective sentences are discarded. In sentence-level analysis sometimes objective sentences have been neglected due to its objective nature. In aspect/word-level analysis a single word is considered for sentiment mining as it is the smallest unit with the specific meaning given in the context. In this type, a lexicon word like an adjective, adverb, verbs are included for classifying the sentiment [9]. Sentiment mining has been performed extracting features for classification. Feature extraction and feature selection are two types of techniques used for features. Feature selection plays an important role by removing unweighted and irrelevant features and improve the overall performance of the model. Feature selection also helps to minimize the model evaluation time. Normally, feature selection techniques have been divided into three categories: filter-based feature selection techniques, wrapper-based feature selection techniques and embedded feature selection techniques. The earlier technique is normally used in data preprocessing. In these types of techniques, features are sorted based on information provided by them before the process of classification. In the second technique, features are selected that are suitable for machine learning algorithms. It works before applying a machine learning algorithm as well as during machine learning algorithms implementation to search out the best suitable feature set. In the third feature selection technique, those features are considered that play an important role during each training iteration of the model and have a noticeable contribution to the training of particular iteration.

The sentiment mining has been performed using various n-gram variations. There are two variations of n-gram word-level and character-level. In the word-level analysis, each word has been considered as a feature that helps in classifying the data according to sentiment [3]. The approaches for sentiment mining classification have been either supervised or unsupervised for machine learning. The other approach is lexicon-based. In the first approach various classifiers have been used includes Naïve Bayes, Logistic Regression, Maximum Entropy, Decision Tree, Random Forest, Support Vector Machine, ID3 etc. and for later approach, the semantic orientation technique has been used widely which helps to identify how near or far that particular term from being positive or negative [9]. Roman Urdu has been widely used for communicating and reviewing. Relational messages and informational messages are two standard distributions for text messaging whereas Roman Urdu has completely different structural and functional properties [13]. Sentiment mining has also been performed using various neural network techniques. These techniques involve convolutional neural network, Long-short term memory and Bi-LSTM. Both n-gram variations have been analyzed on sentence-level and text-level using the Arabic hotel reviews for aspect-based sentiment analysis calculating F1-score and accuracy [21]. Classification of aspect-level sentiment composed of two levels, at first the aspects are extracted by using some specific patterns of different phrases and then determine the polarity employing a rule-based linguistic approach on three various n-gram variations. These are ‘n-gram around’, ‘n-gram after’ and ‘n-gram before’ evaluating on two different datasets one is from computer science and the other one is from bioinformatics [22]. Sentiment analysis for Roman Urdu performed on Daraz.pk reviews. The data is about the reviews of various clothing, electronics and home appliances etc. which ultimately helps to develop a general perception of those particular products. TF-IDF has been used for extracting features and the SVM classifier has been applied for evaluating results [24]. The purpose of using this language is that it is one the most common language used for communication over the digital devices and social applications in South Asian countries and everyone using its own word structure like “2mro” is not a standard spelling of “tomorrow” [34].

Ambiguous Roman Urdu Script:

There are various challenges for Roman Urdu which makes the sentiment mining task more difficult. Some of the challenges are listed below.

- Spell diversion is one of the major issues for Roman Urdu: as there is no standard spelling for a specific word.
    - The word “Khubsurat” {Beautiful} can have multiple spelling styles (‘Khobsurat’, ‘Khobsurat’, ‘Khobsirat’) as per the understanding of the writer.
- In Roman Urdu, a word can behave as multilingual by having more than one meaning with the same spelling.
    - The word “log” {People} might be depicted as “People” whereas the same spelling can be used as a mathematical term “Log”.
- In Roman Urdu there is no capitalization, a single word might be written either with a capital letter or with a small letter using the English alphabets.
    - The name of the place “Lahore” can be written in a sentence with different spellings like “mujhe to murghcholaypasandhainbohatlahore k” {I like chicken chicks very much from Lahore} while “mujhe to murghcholaypasandhainbohat Lahore k”.
- Diacritic spell variation is another issue for Roman Urdu because these diacritics can affect the writing way for a word.
    - A word “ک تاب” {Book} can be written with different variations in Roman Urdu like “Kittab, Kitab, Kattab, Kattaib”.

Dataset

The dataset plays a critical role in classification. Several datasets have been publicly available for sentiment mining classification of resource-rich language. Roman Urdu dataset is a major constraint for classification. We have used the Roman Urdu dataset[1] belongs to six domains. The dataset consists of 11k reviews which are gathered from different websites like “Hamariweb”, “Dramaonline”, “Siasat.pk”, “Whatmobile.com” etc. A detailed description of the dataset is given in table 1. The dataset consists of positive and negative sentences. Those reviews that depict some positive meaning like “Battery timing achihai.speed b axhihai” {Battery timing is good and speed too} is a review about a mobile phone which marked as positive. Whereas those reviews that show some negative meaning like “ye election bohotghalathoayhain aur is dafa ki hukomatkuchnahikarpayegipichlidafa ki hukomat kay banisbat” {These elections have gone very wrong and this time the government will not be able to do anything compared to the previous government} is marked as negative.

Review’s Domain	**Reviews Description**		**Total Reviews**
	Positive	Negative	
Drama/Movie	1335	1208	2543
Politics	646	1835	2481
Sports	238	224	462
Food	658	179	837
Mobile Reviews	693	416	1109
Miscellaneous	2116	1452	3568
Total	5686	5314	11000

Table 1: Descriptive Statistics of Dataset

The table 1 shows that the dataset belongs to six domains in which some categories are clear while there is a category named as miscellaneous in which those reviews are kept which are possessing dual nature. For an example a sentence “Mahira khan ne bohtachi acting ki hailekinagr in ki jagah pe sajjalali is movie me mahira khan ka rooladakarti to kamal tha” {Mahira Khan has acted very well but it would have been great if Sajjal Ali had played the role of Mahira Khan in this movie instead} shows that possess meaning about mahira first that mahira’s acting is good whereas second part gives more positive impact comparatively about sajjalali. In this domain some questionable reviews like “ham kabtakaysa logon ko vote dytyrahyngy?” {How long will we continue to vote for such people?} are also considered. The other five domains have reviews related to the mentioned name of the categories. The details show that the dataset is almost balanced having more positive reviews than negative.

Materials and methods:-

In this section, the detailed methodology has been discussed in which various classifiers have been applied through three feature selection techniques to maximize the results. Initially, the dataset consists of 11k reviews belong to the Roman Urdu domain comprises of positive and negative reviews that have been preprocessed by removing duplications, emojis etc. Our major focus is to analyze the role of feature selection techniques for sentiment mining of the Roman Urdu dataset. Features are selected using different feature selection techniques and experiments are being executed using k-fold cross-validation. We have used 10-fold cross-validation for our experimental setup. Figure 1 presented the detailed methodology steps.

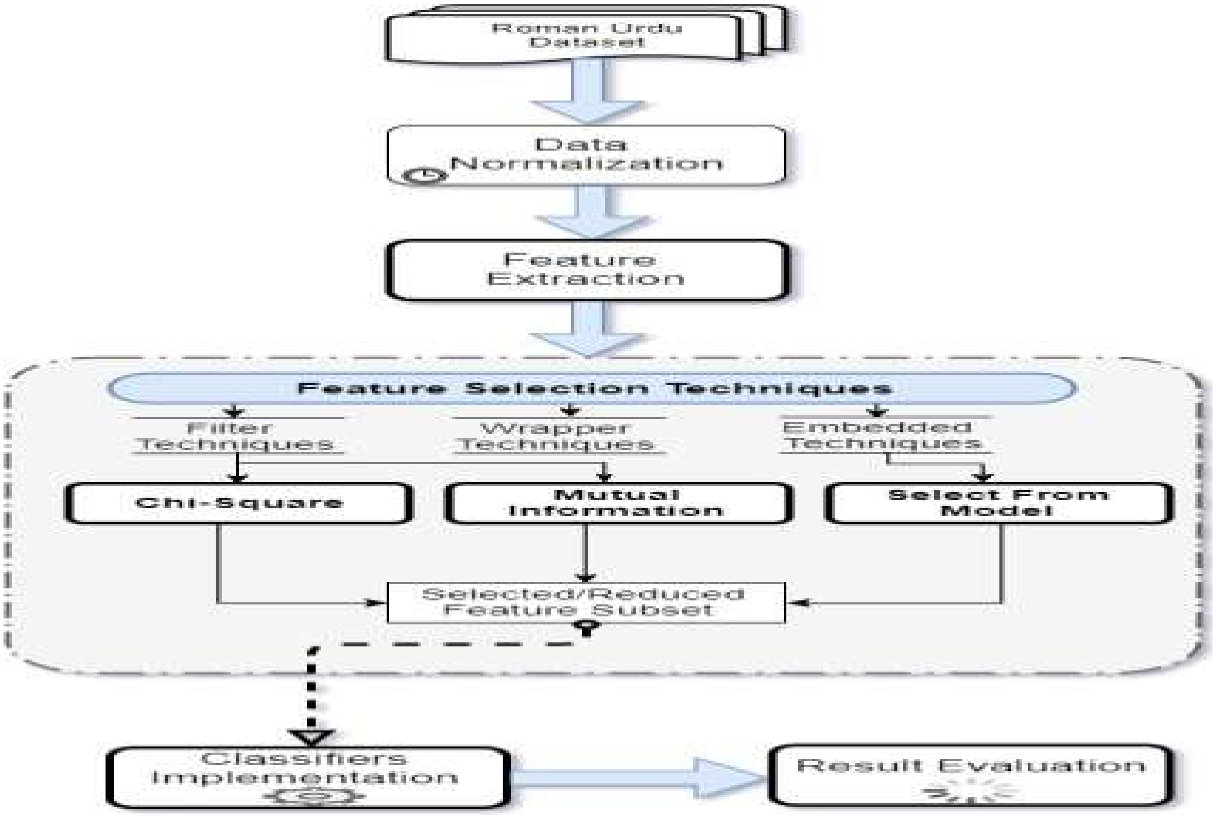


Figure 1:Proposed framework for sentiment classification

Three feature selection techniques have been used for our classification. These techniques help to select those features that show a great impact on sentiment classification. The following techniques have been used for sentiment mining classification.

- Chi-square
- Mutual information
- Select from model

The first technique is a filter-based approach in which the relationship between variables/features has been measured. A specific value χ^2 has been used for measuring.

the relationship in terms of sentiment mining of features. The higher value of χ^2 shows the greater importance of the feature. The following formula in equation 1 has been used for classification in which N is the total number of features and k represents the features selected for classifying the data.

$$(1) \qquad \tilde{\chi}^2(f) = \sum_{k=1}^{C} \frac{n_k}{N} \times \chi^2(f, C_k)$$

The second feature selection technique is mutual information in which a correlation indicates the semantic relation between the features. The higher the correlation value higher will be the dependencies of features on each other. In equation 2 ‘t’ referred to as a term and ‘c’ as a category which helps to analyze the relationship between these two terms.

(2) $$I(t,c) = \log \frac{P_r(t \wedge c)}{P_r(t) * P_r(c)}$$

The third feature selection technique belongs to the embedded category of feature selection which declares a threshold of feature coefficients and only those features are selected that are between the borderline. This task is performed by testing the estimator for those features.

These feature selection techniques have been analyzed on various machine learning classifiers includes logistic regression in which a hypothesis value has been predicted between 0 and 1 using equation 3.

(3) $$0 \le h_\theta(\theta^T x) \le 1$$

The second classifier is Support vector machine that works efficiently for text classification by having an optimized value for training data. The given formula in equation 4 has been used for classification.

(1) $$h_\theta = \begin{cases} 0\ if \theta^T x \le -1 \\ 1\ if \theta^T x \ge 1 \end{cases}$$

We have used linear support vector (svc) variation of SVM which gives efficient results for text type data classification. Multinomial Naïve Bayes is the variation of Naïve Bayes which has been used for our classification problem using equation 5.

(1) $$C_{NB} = argmax_{C \in} C^P(C_j) \prod_i P(x_i \mid y_j)$$

(2)

The other classifier is RandomForest which works by forming the trees for classification. It works for both categorical and numerical features by the meaning of probability calculation according to the specific sentiment class. The Multilayer perceptron is a non-linear uni-directional classifier that includes at least one hidden layer and several non-linear units that effectively show the relation between various inputs. If there is a binary prediction then a single neuron is composed and if there is a non-binary prediction then N number of neurons is composed at an output layer shown in figure 2.

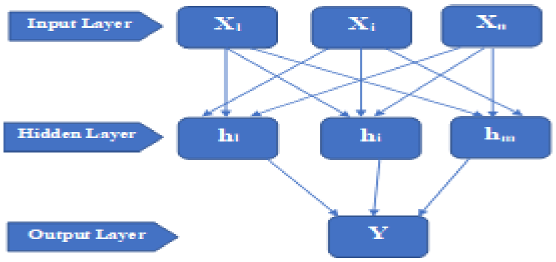


Figure 2:MLP Architecture

The last classifier we have used for our analysis is Decision tree. It works like a flow chart and consists of three types of nodes root nodes, leaf nodes and decision nodes. For classification decision tree calculates the probability according to the given class. It uses a top-down approach and splitting until the values have been assigned to the attributes.

We have analyzed deep neural networks for the classification of sentiments for roman Urdu which shows promising results. We have worked on three neural network techniques for classification which include convolutional neural network, long-short term memory and bi-directional long-short term memory. In the first technique, the focus is on environmental-based learning by storing information in connectional form between the neurons. During training, weights play a vital role by leaning precisely and classify the data according to our label set. Equation 6 has been used for CNN.

(1) $$(f * g)(t) \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} f(\tau) g(t-\tau) d\tau$$

The second technique has four components in which the initial one is a self-connected memory cell whereas three components include input, output and forgets gates. Its core motive is to the classification of data, processing of that data and finally predicting the particular label class according to available classes either positive or negative. The detailed implementation of memory cells is given in equation 7.

$$\begin{aligned} i_t &= \sigma W_i h_{t-1} + U_i x_t + b_i \\ f_t &= \sigma W_f h_{t-1} + U_f x_t + b_f \\ {}^{\sim}\mathrm{c}_t &= tanhW_c h_{t-1} + U_c x_t + b_c \qquad (2) \\ c_t &= f_t \mathit{\Theta} c_{t-1} + i_t \mathit{\Theta}^{\sim}\mathrm{c}_t \\ o_t &= \sigma W_o h_{t-1} + U_o x_t + b_o \\ h_t &= o_t \mathit{\Theta} \tanh{(c_t)} \end{aligned}$$

Where “σ” is used as an element-wise sigmoid function, “$\mathit{\Theta}$” is used element-wise product, “ x_t ” used as input vector, “ h_t ” is the output vector which stores most useful information according to time t. “ U ” describe weight metrics for different gates like input, output etc. according to x_t while on the other hand “ W ” is the weight matrices for hidden state h_t .

Finally, we have analyzed Bi-LSTM which comprises two processing layers first one is a forward layer and the second one is a backward layer. The information regarding a particular feature is stored from left to right sequence and vice versa. In the classification of sentiment mining, it includes the information of a particular feature form both sides and predict the class positive/negative. Network propagation is given in figure 3.

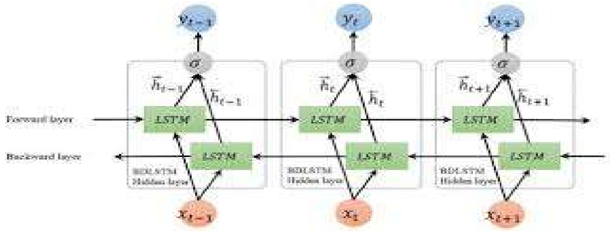


Figure 3:Bi-LSTM Neural Network

Neural network methodology includes the word embedding which is of 300 units and the activation parameter is ‘sigmoid’ whereas ‘Adam’ is an optimizer used for classification. All three variations of neural have been performed on the given parameters.

Results and Discussion:-

We have performed our experiments for classification for two variations of the n-gram. The first one word-level and the second one is character-level. In the earlier variation, we have taken different combinations including Unigram, Bigram, Uni-bi (1-2) gram, Uni-bi-tri (1-3) gram and Uni-bi-tri-four (1-4) gram while in the later variation of n-gram we have experiment variations bigram, trigram, four (4)-gram, five (5)-gram, Six(6)-gram and seven (7)-gram. These variations play an important role in classifying the data according to the sentiment.

The experiments are being carried out using two evaluation measures which include accuracy and F1-score. The accuracy is the ratio of correctly predicted to the total observation. The formula for accuracy is given below.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

In the given equation ‘TP’ Stands for True Positive, ‘TN’ stands for True Negative, ‘FP’ stands for False Positive, and ‘FN’ stands for True Negative.

The F1-score is calculating the mean average of precision and recall by using the formula given below.

$$\text{F1-score} = 2.\frac{Precision\ .Recall}{Precision\ +Recall}$$

In the equation precision is the ratio to find true positive instances to the total number of positively predicted label class while recall is the ratio to correctly classified instances divided by the sum of the number of correctly classified and number of incorrectly classified instances.

Table 2, table 3 and table 4 represents the word-level analysis for all n-gram variations through the various feature set for each feature selection technique. In table 2 chi-square word-level analysis has been performed using the feature set ranges from 3k-14k. The n-gram variations mentioned in the table the complete result analysis in which the best feature set for 1-3 gram is 10k which gives the highest F1-score 89.4% for the Multilayer perceptron classifier. In table 3 mutual information feature selection technique for word-level analysis shows that the feature set ranges from 5k-18k whereas the best result among all the variations for this particular technique calculated for 1-2 gram variation by using 12k feature obtaining the highest F1-score 83.6% for MLP classifier. Table 4 represents the third feature selection technique results for word-level analysis. An estimator has been used for this particular selection technique which gives the best among all feature selection technique result for 1-4 gram variation using 'linear svc' as an estimator which have the highest performing estimator having 90.4% F1-score. This is the maximum result obtained for word-level analysis among all used feature selection techniques.techniques. This particular technique has maximized the result for 4-gram variation of n-gram for character-level analysis.

Table 5, table 6 and table 7 shows the character-level analysis for features selection techniques working with different n-gram variations. In table 5 we have presented the character-level analysis for the chi-square feature selection technique in which the feature set ranges from 1.6k to 8k and 4-gram variation gives the best among all classifiers having 87.5% F1-score for Support vector machine classifier using a feature set of 4k. Table 6 represents the second feature selection technique result for character-level analysis using features ranges from 1.5k-9k. Among all n-gram variations, 4-gram gives the maximum result for SVM classifier that is 84.3% F1-score using 5k features. Table 7 presents the calculated results for the third feature selection technique which gives the best results among all character analyses. It gives the 90.1% F1-score for 4-gram variation of n-gram on SVM machine learning classifier. The estimator 'linear svc' has been used for analyzing n-gram variations for character-level analysis. These variations play an important role in the classification of the data and select from the model technique has outrun the results among all the selection.

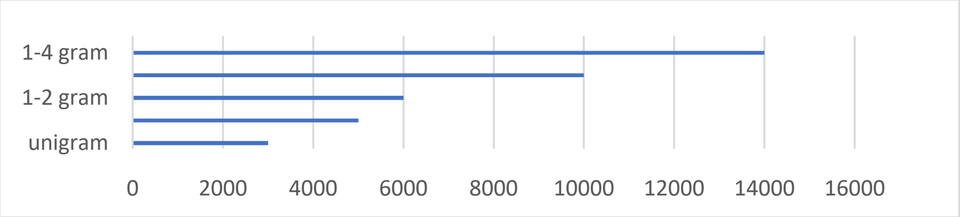
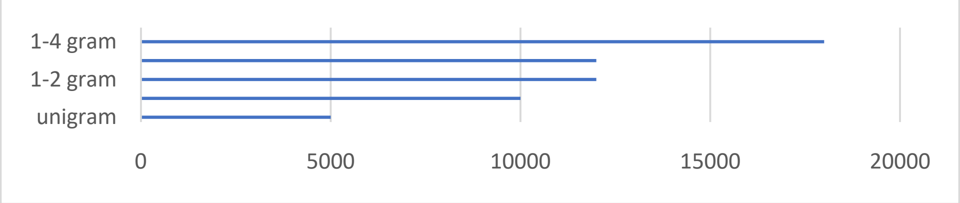
N-gram Value	Unigram		Bigram		1-2 gram		1-3 gram		1-4 gram	
Feature Set for N-gram										
ML classifier	Acc.	F1-score	Acc.	F1-score	Acc.	F1-score	Acc.	F1-score	Acc.	F1-score
LR	0.730	0.835	0.5738	0.715	0.737	0.839	0.742	0.843	0.744	0.844
SVM	0.790	0.875	0.794	0.873	0.814	0.890	0.817	0.892	0.808	0.886
RF	0.790	0.785	0.794	0.773	0.814	0.776	0.817	0.772	0.807	0.772
DT	0.627	0.767	0.659	0.773	0.619	0.757	0.621	0.757	0.621	0.757
MNB	0.712	0.822	0.681	0.801	0.748	0.847	0.753	0.850	0.746	0.846
MLP	0.789	0.874	0.817	0.889	0.805	0.884	0.822	**0.894**	0.807	0.885

Table 2: Chi-Square Word-Level ML ClassifiersResults

N-gram Value	Unigram		Bigram		1-2 gram		1-3 gram		1-4 gram	
Feature Set for N-gram										
ML classifier	Acc.	F1-score	Acc.	F1-score	Acc.	F1-score	Acc.	F1-score	Acc.	F1-score
LR	0.663	0.786	0.501	0.633	0.692	0.808	0.683	0.801	0.684	0.802
SVM	0.701	0.815	0.638	0.762	0.731	0.835	0.717	0.825	0.725	0.831
RF	0.701	0.752	0.638	0.732	0.731	0.767	0.717	0.768	0.725	0.763

DT	0.606	0.751	0.627	0.742	0.614	0.753	0.617	0.752	0.614	0.746
MNB	0.589	0.729	0.532	0.682	0.643	0.772	0.624	0.752	0.615	0.746
MLP	0.709	0.821	0.516	0.627	0.733	**0.836**	0.68	0.793	0.576	0.661

Table 3:Mutual Information Word-Level ML Classifiers Results

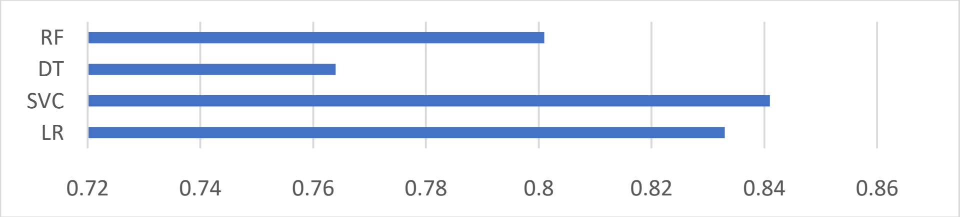
N-gram Value	**Unigram**		**Bigram**		**1-2 gram**		**1-3 gram**		**1-4 gram**	
Estimator										
ML classifier	Acc.	F1-score	Acc.	F1-score	Acc.	F1-score	Acc.	F1-score	Acc.	F1-score
LR	0.757	0.853	0.589	0.731	0.761	0.854	0.758	0.853	0.759	0.855
SVM	0.825	0.897	0.773	0.861	0.836	0.904	0.833	0.902	0.837	**0.904**
RF	0.825	0.792	0.773	0.791	0.836	0.786	0.833	0.789	0.837	0.785
DT	0.658	0.783	0.656	0.773	0.639	0.767	0.645	0.771	0.633	0.766
MNB	0.73	0.835	0.644	0.773	0.751	0.849	0.751	0.848	0.751	0.849
MLP	0.812	0.888	0.768	0.858	0.728	0.795	0.821	0.894	0.823	0.896

Table 4:Select from Model Word -Level ML Classifiers Results

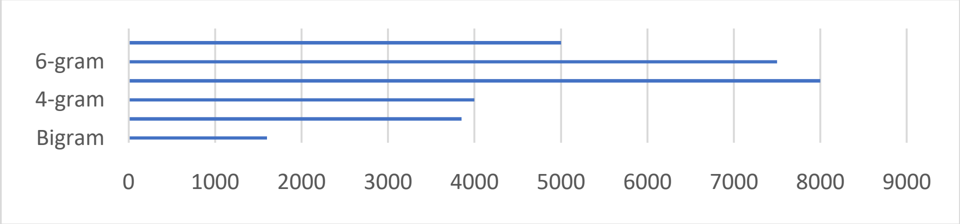
N-gram Value	**Bigram**		**Trigram**		**4-gram**		**5-gram**		**6-gram**		**7-gram**	
Feature Set for N-gram												
ML classifier	Acc.	F1-score	Acc.	F1-score	Acc.	F1-score	Acc.	F1-score	Acc.	F1-score	Acc.	F1-score
LR	0.681	0.799	0.741	0.842	0.75	0.848	0.754	0.851	0.756	0.845	0.692	0.808
SVM	68.53	0.803	0.765	0.858	0.796	0.879	0.795	0.878	0.791	**0.875**	0.691	0.874
RF	68.53	0.724	0.765	0.764	0.794	0.822	0.795	0.834	0.792	0.833	0.791	0.838
DT	58.76	0.728	0.626	0.766	0.667	0.789	0.691	0.805	0.695	0.811	0.708	0.818
MNB	50.92	0.661	0.697	0.811	0.746	0.845	0.759	0.854	0.763	0.856	0.736	0.839
MLP	66.57	0.789	0.735	0.838	0.784	0.871	0.788	0.872	0.779	0.867	0.768	0.861

Table 5:Chi-Square Character Level ML Classifiers Results

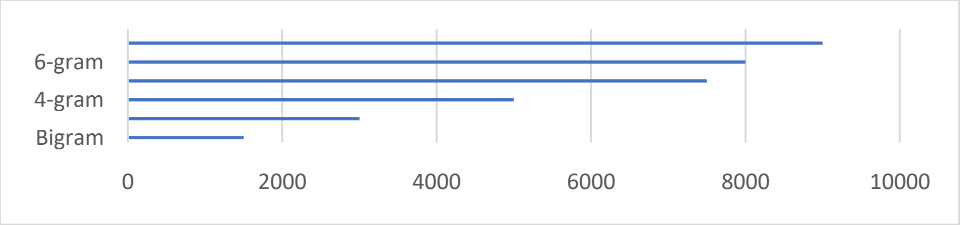
N-gram Value	Bigram		Trigram		4-gram		5-gram		6-gram		7-gram	
Feature Set for N-gram	 											
ML classifier	Acc.	F1-score	Acc.	F1-score	Acc.	F1-score	Acc.	F1-score	Acc.	F1-score	Acc.	F1-score
LR	0.681	0.799	0.702	0.815	0.723	0.829	0.704	0.816	0.688	0.805	0.658	0.783
SVM	0.686	0.803	0.731	0.835	0.742	**0.843**	0.725	0.831	0.711	0.821	0.683	0.802
RF	0.686	0.719	0.731	0.761	0.742	0.811	0.725	0.824	0.711	0.813	0.683	0.796
DT	0.589	0.731	0.629	0.761	0.671	0.766	0.675	0.798	0.677	0.794	0.674	0.792
MNB	0.511	0.663	0.602	0.741	0.703	0.815	0.705	0.816	0.701	0.814	0.694	0.809
MLP	0.663	0.784	0.717	0.825	0.709	0.819	0.725	0.824	0.711	0.815	0.635	0.761

Table 6: Mutual Information Character levelML Classifiers Results

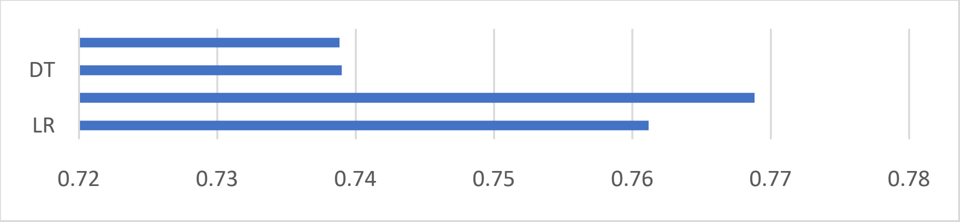
N-gram Value	Bigram		Trigram		4-gram		5-gram		6-gram		7-gram	
Estimator	 											
ML classifier	Acc.	F1-score	Acc.	F1-score	Acc.	F1-score	Acc.	F1-score	Acc.	F1-score	Acc.	F1-score
LR	0.695	0.81	0.762	0.856	0.789	0.874	0.774	0.865	0.759	0.855	0.731	0.835
SVM	0.717	0.826	0.804	0.884	0.839	0.905	0.832	**0.901**	0.822	0.896	0.789	0.874
RF	0.717	0.734	0.804	0.779	0.839	0.829	0.832	0.845	0.822	0.837	0.789	0.836
DT	0.591	0.728	0.642	0.766	0.688	0.801	0.707	0.816	0.704	0.814	0.703	0.815
MNB	0.494	0.647	0.711	0.821	0.763	0.857	0.761	0.856	0.757	0.853	0.741	0.842
MLP	0.705	0.817	0.695	0.774	0.811	0.889	0.814	0.89	0.784	0.869	0.712	0.775

Table 7: Select from model Character level ML Classifiers Results

Algorithm	Acc.	F1-score
CNN	0.917	0.915
LSTM	0.908	0.903
Bi-LSTM	0.918	**0.917**

Table 8: Neural Network Classification for Roman Urdu

Table 8 presents the neural network analysis result for the sentiment mining classification of Roman Urdu data. Three variations of deep learning variations have been applied in which the highest F1-Score achieved by Bi-LSTM 91.7%. LSTM and CNN achieved slightly fewer results than Bi-LSTM. These results are obtained by employing word embedding. In which 300-dimensional embedding has been performed and Bi-LSTM classifies the data through the forward and backward knowledge which effectively classifies the data.

We have compared the proposed results with the baseline results presented by [3] in which they have calculated the accuracy for both variations of n-gram for Roman Urdu sentiment mining. Figure 4 represents the detailed comparison between the results. In terms of word-level analysis, we have obtained 83.7% compared to 80.8% baseline results. We have also surpassed the character-level result in which we have achieved 83.9% in comparison to 81.2%. These results are 2.84% and 2.76% improved respectively. Deep learning analysis outperformed the baseline deep learning. We have achieved 91. 8% accuracy in contrast to 78.2% baseline accuracy which is 13.46% better than the previously reported one. We have also represented the overall best result which is 9.3% higher than the baseline.

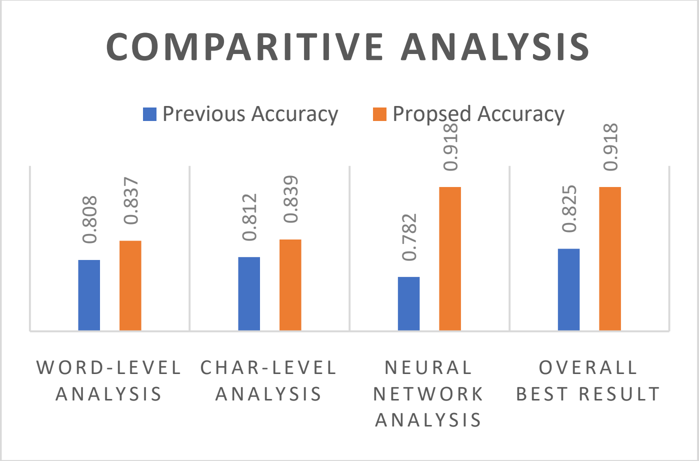


Figure 4:Comparative Analysis with baseline results

Sentiment mining classification has always been a major challenge due to the unavailability of resources and labeled data. We have performed this task using various classification strategies. Feature selection for sentiment mining for Roman Urdu has emerged as a potential task to classify the data according to the context. We have analyzed different n-gram variations due to a large number of linguistic features. Various feature selection techniques have been applied so that the quality feature set has been incorporated. In the case of word-level feature analysis, it is observed that two words ‘chor’ {Thief} and ‘shor’ {Noise} belong to the same class whereas in the case of character-level analysis some parts are the same some are different. Considering an example for bi-gram character-level analysis “-c,ch,ho,or,r-” & “-s,sh,ho,or,r-” three bigrams are same and two are different. These analysis affects the overall classification of the dataset and helps to correctly classify in the given context.

Feature selection techniques give promising results in terms of classification because it helps to reduce the size of the feature set by removing unwanted and unweighted features. These techniques also help to avoid overfitting by ignoring duplicate data. Chi-square feature selection technique has been implemented; a deep analysis shows that those features are weighted more that are directly dependent on each other. In the case of word-level analysis, a bigram word “kehta ho” {Say} feature has more weight than “main itni” {I so} feature bigram. The first bigram doesn’t give any information individually while in the second bigram both words have some independent information. Mutual information is the secondly implemented feature selection technique for both n-gram variations. Critical analysis of results shows that more valued features are those that are less occurred and don’t depict any clear information given the least importance. For example, a bigram feature in word-level analysis “aur ghatiya” {And crap} has a more weighted feature than “hai us” {Is that} bigram feature. Among the selection techniques we have applied in the first technique those features are of low value that is independent to some extent while in the second technique low valued features are those that are not giving any information about the other feature. The last feature selection technique is select from a model that works based on estimators. We have analyzed multiple machine learning classifiers to get the best-suited estimator. It is being observed that linear svc gives the most promising results. In this technique estimation employing coefficient, the estimator selects the more weighted features similar to previous selection techniques. In the features list, “aajtak” {Till today} have a more weighted feature than “ho skty” {May be} feature bigram fir word-level analysis. These techniques help to minimize the feature set size by selecting the best suitable feature for all the variations. We have also achieved the best results on our dataset through discussed feature selectin techniques than the reported baseline results presented by [2]. Convolutional neural network and Recurrent neural network includes LSTM, Bi-LSTM has shown great achievement in recent times for text and speech type of data. We have analyzed that among these techniques Bi-LSTM gives efficient results. The reason being showing promising results is the pre-trained word embedding which helps to classify the data in the context of the actual meaning of the sentence. The critical analysis of neural network for sentiment mining shows that the forward and backward information plays an integral part to correctly classify the data. The embedded dimensions and optimizer give the maximum results.

Conclusion:-

Classification for Roman Urdu sentiments plays an important role because of a large amount of online unorganized data. This helps to extract valuable decision-making data after classification. Different feature selection and feature extraction techniques being used for the classification of data according to sentiments. The proposed methodology comprises three feature selection techniques that are implemented on the Roman Urdu dataset for sentiment classification. These techniques are chi-square, Mutual Information and select from model which are experimented using two variations of n-gram i-e word-level and character-level. The core motive behind these variants' classification is to dig out the best possible combination. Six machine learning classifiers including LR, SVM, DT, RF, MNB, MLP and three neural network techniques namely CNN, LSTM and Bi-LSTM which are evaluated using Accuracy and F1-Score evaluation measures. In terms of n-gram variations for word-level are Unigram, Bigram, Uni-Bi gram, Uni-Bi-Tri gram and Uni-Bi-Tri-Four gram and for character-level analysis, we have used n-gram from bigram to 7-gram.

The results show an encouraging improvement from the baseline results in terms of word-level and character-level classification. We have achieved 91.8% accuracy for Bi-LSTM that is the highest one score among all the results which is 9.3% in contrast to previous work. In the word-level analysis, we evaluated 2.9% better results than baseline and for character-level analysis, we got 2.7% better results. We have a noticeable improvement in neural network results in which we achieved 13.6% accuracy from baseline. These results show the effectiveness of feature selection for sentiment mining classification for Roman Urdu. The highest F1-Score is 91.7% that is achieved for Bi-LSTM. There will be a potential path to employ a feature selection technique with a neural network to enhance the performance for the classification of sentiment mining Roman Urdu. We will work for rulemaking of the Roman Urdu text to apply classification strategies on it.

References:-

(1) A. Alam and J. Hussain, "EnSWF: effective features extraction and selection in conjunction with ensemble learning methods for document," Applied Intelligence, vol. 49, no.8, pp.3123-3145, 2019.

(2) K. Mehmood, D. Essam, and K. Sha, "Sentiment Analysis System for Roman Urdu," Science and Information Conference, Springer, pp. 29–42, 2019.

(3) K. Mehmood, D. Essam, and K. Shafi, "Sentiment Analysis for a Resource-Poor Language — Roman Urdu," ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), vol. 19, no. 1, pp. 1–15, 2019.

(4) A. Abbasi, S. France, and H. Chen, "Selecting Attributes for Sentiment Classification using Feature Relation Networks", IEEE Transactions on Knowledge and Data Engineering, no. 3, pp. 447-462, 2011.

(5) K. Mehmood, D. Essam, K. Shafi, and M. K. Malik, "Discriminative Feature Spamming Technique for Roman Urdu Sentiment Analysis," IEEE Access, no. 7, pp. 47991-48002, 2019.

(6) M. T. Khan, M. Durrani, A. Ali, I. Inayat, and S. Khalid, "Sentiment analysis and the complex natural language," Complex Adapt. Syst. Model., vol. 4, no. 1, pp. 1-19, 2016.

(7) B. Tina, "Urdu – Roman Transliteration via Finite State Transducers," In FSMNLP 2012, 10th International Workshop on Finite State Methods and Natural Language Processing, pp. 25–29, 2012.

(8) Z. Sharf and S. U. Rahman, "Lexical normalization of roman Urdu text," International Journal of Computer Science and Network Security, vol. 17, no. 12, pp. 213–221, 2017.

(9) A. Rexha and M. Kr, "Opinion Mining with a Clause-Based Approach," In Semantic Web Evaluation Challenge Springer, vol. 2, pp. 166–175, 2017.

(10) B. Saberi and S. Saad, "Sentiment Analysis or Opinion Mining: A Review Sentiment Analysis or Opinion Mining: A Review," IJASEIT, no.5, 2018.

(11) Q. Liu, Z. Gao, B. Liu, and Y. Zhang, "Automated Rule Selection for Aspect Extraction in Opinion Mining," Ijcai, pp. 1291–1297, 2015.

(12) S. Khalid, M H. Aslam, and M. T. Khan. "Opinion Reason Mining: Implicit Aspects beyond Implying aspects." In 2018 21st Saudi Computer Society National Computer Conference (NCC), pp. 1-5. IEEE, 2018.

(13) M. Bilal, H. Israr, M. Shahid, and A. Khan, "Sentiment classification of Roman-Urdu opinions using Naive Bayesian, Decision Tree and KNN classification techniques," J. King Saud Univ. - Comput. Inf. Sci., vol. 28, no. 3, pp. 330–344, 2016.

(14) A. Bilal and A. Kakakhel, "Roman-txt: Forms and Functions of Roman Urdu Texting," In Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services, pp 1-9, 2017.

(15) A. Z. Syed and A. Nazir, “Mining the Urdu Language-Based Web Content for Opinion Extraction,” In Mexican Conference on Pattern Recognition, vol. 3, pp. 244–253, 2017.

(16) K. Khan and A. Ur, “Urdu Sentiment Analysis,” (IJACSA) International Journal of Advanced Computer Science and Applications, vol. 9, no. 9, pp. 646–651, 2018.

(17) M. Hassan and M. Shoaib, “Opinion within Opinion: Segmentation Approach for Urdu Sentiment Analysis,” Int. Arab J. Inf. Technol., vol. 15, no. 1, pp. 21–28, 2018.

(18) S. Almatarneh, “Comparing Supervised Machine Learning Strategies and Linguistic Features to Search for Very Negative Opinions,” Information, vol. 10, no.1, pp. 16,2019.

(19) G. Badaro, R. Baly, and H. Hajj, “A Large-Scale Arabic Sentiment Lexicon for Arabic Opinion Mining,” In Proceedings of the EMNLP 2014 workshop on Arabic natural language processing (ANLP), pp. 165–173, 2014.

(20) L. Yu, L. Lee, J. Wang, and K. Wong, “IJCNLP-2017 Task 2: Dimensional Sentiment Analysis for Chinese Phrases,” In Proceedings of the IJCNLP, Shared Tasks, pp. 9–16, 2017.

(21) M. Al-smadi, “Using long short - term memory deep neural networks for aspect-based sentiment analysis of Arabic reviews,” Int. J. Mach. Learn. Cybern., vol. 10, no. 8, pp. 2163-2175, 2018.

(22) M. Touseef, I. Muhammad, and T. Afzal, “Aspect based citation sentiment analysis using linguistic patterns for better comprehension of scientific knowledge,” Scientometrics, vol. 119, no. 1, pp. 73-95, 2019.

(23) J. Im, T. Song, Y. Lee, and J. Kim, “Confirmatory Aspect-based Opinion Mining Processes.” arXiv preprint arXiv:1907.12850, 2019.

(24) S. Schmunk, W. Höpken, M. Fuchs, and M. Lexhagen, “Sentiment Analysis: Extracting Decision-Relevant Knowledge from UGC.” In Information and Communication Technologies in Tourism, Springer, pp. 253-265, 2013.

(25) F. Noor, M. B. B, and J. Baber, “Sentiment Analysis in E-commerce Using SVM on Roman Urdu Text.” In International Conference for Emerging Technologies in Computing, Springer, pp. 213-222, 2019.

(26) B. Agarwal and N. Mittal, “Machine Learning Approach for Sentiment.” In Prominent feature extraction for sentiment analysis, Springer, pp. 21-45. 2016

(27) R. Islam and M. F. Zibran, “SentiStrength-SE: Exploiting domain-specific city for improved sentiment analysis in software engineering text,” Journal of Systems and Software, vol. 145, pp. 125–146, 2018.

(28) J. Pre-proofs, M. Asif, A. Ishtiaq, H. Ahmad, H. Aljuaid, and J. Shah, “Sentiment Analysis of Extremism in Social Media from Textual,” Telematics and Informatics, no. 48, 2020.

(29) G. Bologna, “applied sciences A Simple Convolutional Neural Network with Rule Extraction,” Applied Sciences, vol. 9, no.12, pp. 2411, 2019.

(30) H. T. Phan and V. A. N. C. Tran, “Improving the Performance of Sentiment Analysis of Tweets Containing Fuzzy Sentiment Using the Feature Ensemble Model,” IEEE Access, vol. 8, pp. 14630-14641, 2020.

(31) N. Mukhtar and M. Abid, “Effective lexicon-based approach for Urdu sentiment analysis,” Artif. Intell. Rev., pp. 1-28, 2019.

(32) C. Pong-inwong, “Improved Sentiment Analysis for Teaching Evaluation Using Feature Selection and voting ensemble learning integration,” 2nd IEEE international conference on computer and communications (ICCC)., pp. 1222–1225, 2016.

(33) A. Dogan, “A Weighted Majority Voting Ensemble Approach for Classification,” 4th International Conference on Computer Science and Engineering (UBMK)., pp. 1-6, 2019.

(34) F. H. Khan, U. Qamar, and S. Bashir, “SentiMI : Introducing Point-wise Mutual Information with SentiWordNet to Improve Sentiment Polarity Detection SentiMI : Introducing point-wise mutual information with SentiWordNet to improve sentiment polarity detection,” Appl. Soft Comput. J., vol. 39, pp. 140–153, 2016.

(35) M. A. Sanchez-perez, I. Markov, and G. Helena, “Comparison of Character n-grams and Lexical Features on Author, Gender , and Language Variety Identification on the Same Spanish News Corpus (preprint version) Comparison of Character N-grams and Lexical Features on Author , Gender , and Language Variety Identification on the Same Spanish News Corpus Preprint version,” International Conference of the Cross-Language Evaluation Forum for European Languages, Springer., pp. 145-151, 2017.

(36) M. Ahmad, S. Aftab, S. S. Muhammad, and S. Ahmad, “Machine Learning Techniques for Sentiment Analysis: A Review,” Int. J. Multidiscip. Sci. Eng, vol. 8, no. 3, pp. 27, 2017.

(37) Pingle, A., Vyawahare, A., Joshi, I., Tangsali, R., Kale, G., & Joshi, R. “Robust Sentiment Analysis for Low Resource languages Using Data Augmentation Approaches: A Case Study in Marathi.”2023.

(38) Soomro, M. A., Memon, R. N., Chandio, A. A., Leghari, M., & Soomro, M. H. “A dataset of Roman Urdu text with spelling variations for sentence level sentiment analysis” Data in Brief, 57, 111170, 2024.

(39) Samreen A, Ali SA. Dataset construction to detect human behavior with the help of emotions, sentiments and mood for Roman Urdu. Data in Brief, vol. 52, Pages 100626, 2023.

(40) Koto, Fajri; Beck, Tilman; Talat, Zeerak; Gurevych, Iryna; Baldwin, Timothy. “Zero-shot Sentiment Analysis in Low-Resource Languages Using a Multilingual Sentiment Lexicon.” Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Volume 1: Long Papers, pp. 298-320, 2024.